Award Number: *W81XWH-07-2-0112*

TITLE: *Literature Mining of Pathogenesis-Related Proteins in Human Pathogens for Database Annotation*

PRINCIPAL INVESTIGATOR: *Cathy H. Wu, Ph.D.*

CONTRACTING ORGANIZATION:

Georgetown University Medical Center
Washington, DC 20007

REPORT DATE: *October 2009*

TYPE OF REPORT: *Final*

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From - To)*
01-10-2009	Final	20 Sep 2007 – 19 Sep 2009

4. TITLE AND SUBTITLE

Literature Mining of Pathogenesis-Related Proteins in Human Pathogens for Database Annotation

5a. CONTRACT NUMBER

5b. GRANT NUMBER
W81XWH-07-2-0112

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Cathy H. Wu, Ph.D.
wuc@georgetown.edu

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Georgetown University Medical Center

Washington, DC 20007

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

US Army Medical Research and Material Command
Fort Detrick, Maryland 21702-5012

10. SPONSOR/MONITOR'S ACRONYM(S)

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Biomedical literature represents the primary source of experimental data and biological knowledge. This project developed a text mining system for pathogens of biodefense relevance, focusing on mining pathogen-host proteomic data. We developed a Support Vector Machine (SVM)-based system to identify abstracts containing protein interaction information using an annotated corpus of 1360 MEDLINE abstracts as the training set. It achieved good performance on document classification with a precision of over 80% among top 50 ranked abstracts. The SVM-based method is further augmented with other text mining tools (such as PIE) for mining and tagging PPI information. As part of an effort in enabling text mining tools for real-world applications, we coupled our analysis with the functional annotation of proteomic experiment. All the data was then loaded into iProXpress system and provided to the collaborating USAMRIID laboratory for the analysis of bacterial pathogen proteomics data.

15. SUBJECT TERMS

Text mining, pathogenesis, pathogen-host protein-protein interaction, literature corpus, machine learning, support vector machine (SVM), proteomics data analysis

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			**19b. TELEPHONE NUMBER** *(include area code)*
U	U	U	UU	82	

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39.18

Table of Contents

INTRODUCTION:

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there are growing interests and pressing needs in speeding up the basic research as well as data mining of pathogenesis-related proteins in pathogens of military relevance, which may lead to better targets for disease diagnosis, prevention and therapy. This project specifically focused on pathogenesis-related protein data mining from scientific literature by developing an automated text mining system to facilitate literature-based curation of such proteins (1st year), and from proteomics and functional genomics data through an integrated protein bioinformatics analysis system (2nd year, revised). We refer to the project as the **Pathogen Mining System**. The text mining system development primarily concerns the pathogen-host protein-protein interaction (PH-PPI) information from MEDLINE abstracts. The proteomics and genomics data mining concerns the analysis of proteomics data from *Burkholderia* under simulated growth condition, a project under the **Cooperative Research and Development Agreement with USAMRIID** (USAMRMC Control No: W81XWH-09-0003) [**Appendix I**].

BODY:

YEAR ONE

The primary objective of the first year of the project was to develop a text-mining system to identify pathogenesis-related papers and extract information on pathogenicity and host-pathogen interactions. There were three tasks:

- ***Task1 (M01-03): Compilation of training and benchmarking literature corpus.*** Manual compilation of literature corpus as a positive training set of 300 pathogenesis-related papers with pathogen-host protein-protein interaction information.
- ***Task2 (M04-09): Development and evaluation of text-mining algorithms.*** Development of a text-mining system for document retrieval, entity recognition, and document categorization. Named-entity tagging tools as well as algorithms for document classification and information extraction, including machine learning and rule-based methods evaluated.
- ***Task3 (M10-12): Development of web interface for automated literature mining.*** Development of web-based graphical user interface for query submission and for literature mining result display with automatically tagged abstracts.

I. Literature data sets for machine learning algorithm training

Literature data sets (literature corpus) consisting of positive and negative data are necessary for training machine learning algorithms, such as Supporter Vector Machine (SVM), for text mining of pathogenesis-related pathogen and host proteins from literature. We focused on specific pathogen and host protein-protein interactions (PH-PPI). Unlike those for protein-protein interactions of the same species taking place within an organism, curated positive training data sets are rare for PH-PPI, especially for bacterial PH-PPI, and most such data are buried in the literature. Also because the bacterial PH-PPI information is much more difficult to distinguish from the same-species PPI than viral PH-PPI information would, we decided to separate training

set for the bacterial PH-PPI from that of viral PH-PPI, and to concentrate on the former. Thus, we generated the literature training sets through manual curation of a set of ~2000 abstracts retrieved from PubMed based on query terms "bacterial pathogen and protein interaction".

1. Positive literature set of PH-PPIs. We compiled 300 abstracts (PMIDs) that are reviewed to contain PH-PPI, and the sentences providing the evidence for such interactions are also tagged (highlighted). The sources for deriving the set of literature also include protein database s (UniProtKB and IntAct) where literature with protein interactions is cited for protein entries. Of the 300 abstracts, ~54% are for viral-host PPI, which are all derived from literature cited in databases; while ~46% are for bacterial-host PPI, most of which are from PubMed search. Because the primary interests of pathogens for the USAMRIID are on CDC category A/B viral and bacterial pathogens, the abstracts for training have a balanced coverage of the bacterial and viral groups of organisms. In the training set, viral pathogens include Ebola, Lassa, HIV, HBV and bacterial pathogens include *Yersinia pestis*, *Bacillus anthracis*, *Salmonella*, and *Shigella*. In most cases the host is human, but may also include other mammal species.

2. Negative literature set of PH-PPIs. Of the ~2000 abstracts retrieved from PubMed based on general keyword search "bacterial host protein interaction", ~1225 abstracts were manually selected as negative ones, which may describe pathogen gene- or protein-related information but clearly lack of specific PH-PPI information.

The data sets for bacterial PH-PPI are available at http://pir.georgetown.edu/staff/huz/tatrc/ (*tatrc_dataset_positive.html* and *tatrc_dataset_negative.html*), including 135 positive and 1225 negative abstracts. Evidence sentences in the positive abstracts were also annotated. The data set is currently for internal use and will eventually be made public for use in developing text mining algorithms by the text mining community.

II. Machine learning algorithm development for text mining of pathogenesis proteins

We developed and evaluated machine learning-based text-mining methods for retrieving MEDLINE abstracts containing pathogen and host protein-protein interaction information based on the literature training set. We used a publicly available Support Vector Machine (SVM) package, SVM*light* (see http://svmlight.joachims.org/), to train the classifier, and tested and evaluated both abstract- and sentence-based classifiers to recognize PH-PPI-containing abstracts. Detailed methodology and results are described in a conference paper (Xu *et al*., 2008) [**Appendix II**] and a journal article (Yin *et al*., 2009) [**Appendix III**].

1. Abstract-based algorithm. The training task can be at abstract level (ALT) to build a system to rank a set of abstracts. The abstracts in the dataset were preprocessed first by normalizing the nouns, verbs, and adjectives, followed by extracting the unigrams and bigrams in both title and abstract to construct the sample features. The SVM was trained to classify these 1360 abstracts (both positive and negative) by 10-fold cross-validation. Given a threshold value, abstracts with scores higher than the threshold from the classifier were assigned positive, while those with lower scores labeled negative. The classification was based on the total feature of the abstract. We tried different kernel functions in SVM including linear function, polynomial, and RBF and found linear function was the best.

2. Sentence-based algorithm. The training task can also be at sentence level (SLT) to build a system to rank the abstracts. Individual sentences from abstracts were first extracted and labeled with corresponding PubMed ID (PMID) appended with a sequential number of the sentence in the given abstract. The sentences were then preprocessed similarly as above in the abstract-based algorithm. Untagged sentences from positive abstracts were not used for training but included in the test dataset only. The SVM was trained with linear function at the sentence-based, and 10-fold cross-validation was used to construct training and test dataset. Each sentence received a score from the classifier, and the highest sentence score would be assigned to the abstract as the final discriminating value. Similar to ALT method, a threshold value was set to assign positive or negative abstracts from the classifier, but the classification in SLT method was based on the feature of sentences.

3. Results and comparison between ALT and SLT methods. The testing results of the trained SVM were evaluated using the ROC curved epicting the relationship between the true positive (TP) and false positive (FP) rates (**Figure 1**). In the high specificity area (specificity=1-FP, towards the left of the ROC curve), given the same sensitivity (TP), the sentence-based method gave higher specificity (red-line) than the abstract-based (blue-line); while in the high sensitivity area (sensitivity=TP, towards the top of the ROC curve), the two methods seemed to have little difference. For example, the top 200-scored abstracts from the classifier using sentence-based method contained 61% true positive abstracts, compared to 53% with abstract-based method. The results suggest that the sentence-based training method tends to have better performance than the abstract-based method for retrieving pathogen host PPI abstracts. We also extended the SVM training to feature selection to enhance its performance.

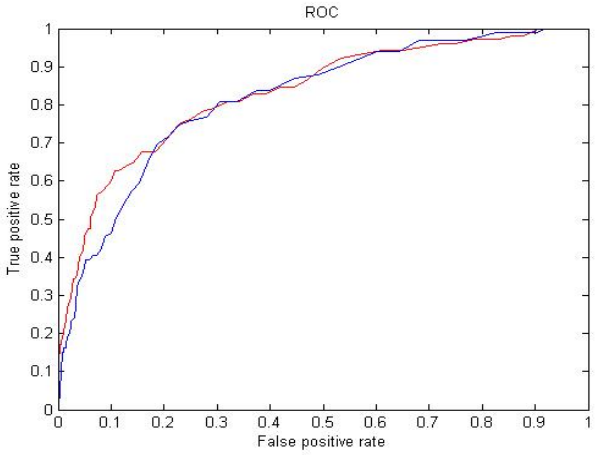


Figure 1. Receiver operating characteristics curve (ROC) analysis of ALT (blue) and SLT (red).

4. Feature selection method and information gain. We investigated the inclusion of a feature selection method (i.e., information gain) into the machine leaning system. We compared *no feature selection* method with *Information Gain* feature selection on both abstract and sentence levels. We found that *Information Gain* reduced the dimension of Vector Space and could improve the performance of the SVM than *no feature selection*. Moreover, the results showed that the sentence-level SVM (training based on highlighted sentences) had better performance and greater prospect than the abstract-based method.

III. Evaluation of existing text mining tools on the PH-PPI data sets

While developing and evaluating the SVM-based text mining system for PH-PPI during the first year of the project, we are also exploring the existing text mining tools that can be useful for text

mining of PH-PPI information. These public text mining tools include PIE (Kim et al., 2008), iHOP (Fernández et al., 2007), and others as included in MetaServer (Leitner et al., 2008), which is a central sever integrating text mining tools participating in the BioCreative Challenge Evaluation for molecular (gene and protein) data from literature (Hirschman et al., 2005). Protein-protein interaction text mining has been a major task in the 2nd BioCreative Challenge Evaluation (Wilbur et al., 2007).

We evaluated the PPI text mining tool PIE (Protein Interaction information Extraction, http://pie.snu.ac.kr/index.php) using the curated positive data set for bacterial as well as the viral PH-PPI. PIE highlights sentences in abstracts that contain protein interaction information, in which the detected words/phrases for the interacting proteins and the interaction relations are also distinguished. **Table 1** (bacterial set) and **Table 2** (viral set) summarize the comparison of the PIE PPI extraction with the manual annotated abstracts and sentences.

Table 1. Comparison of PIE text mining of PPI to the manual bacterial data set

		# Abstracts	% Data set
Abstract level	Manually-tagged bacteria data set	135	100%
	Positive abstracts tagged by PIE	110	**81.5%**
	Positive abstracts not tagged by PIE	25	18.5%
	Abstracts with >=1 manually-identified sentence tagged by PIE	70	51.9%
	Abstracts with no manually-identified sentence tagged by PIE	65	48.1%
Sentence level	Manually-tagged (positive) sentences in data set	247	100%
	Positive sentences tagged by PIE	98	39.7%
	Positive sentences missed by PIE	149	60.3%
	Sentences tagged by PIE in data set	298	100%
	Positive sentences tagged by PIE	98	32.9%
	Negative sentences tagged by PIE	200	67.1%

Table 2. Comparison of PIE text mining of PPI to the manual viral data set

		# Abstracts	% Data set
Abstract level	Manually-tagged virus data set	170	100%
	Positive abstracts tagged by PIE	163	**95.9%**
	Positive abstracts not tagged by PIE	7	4.1%
	Abstracts with >=1 manually-identified sentence tagged by PIE	145	85.3%
	Abstracts with no manually-identified sentence tagged by PIE	25	14.7%
Sentence level	Manually-tagged sentences (positive) in the data set	279	100%
	Positive sentences tagged by PIE	205	**73.5%**
	Positive sentences missed by PIE	74	26.5%

The results show that PIE recognizes ~82% of the manually tagged abstracts and ~40% manually tagged sentences for the bacterial data set, and recognizes ~96% manually tagged abstracts and 74% manually tagged sentences for the viral data set. While we need to compare the PIE's performance with other similar tools on the same data set, the relatively high recognition of

positive abstracts by PIE is a desired feature for retrieving the PH-PPI containing abstracts to facilitate the manual curation efforts. Therefore the PIE tool can augment the pathogen mining system for this project. The detailed evaluation results of the PIE tool are available at: http://pir.georgetown.edu/staff/huz/tatrc/dataset/ with the bacterial set (PIE_evaluation_bacterial_positive.mht) and the viral set (PIE_evaluation_viral_positive.mht).

IV. iProLINK framework to link text mining to ontology and systems biology

Another ongoing effort relevant to the project on the PH-PPI text mining is the iProLINK framework development, an effort in bringing together text mining, biological ontology and systems biology communities to develop text mining tools that can be broadly utilized by the biology communities for real-world applications.

The ever-increasing scientific literature and the exponential growth of large-scale molecular data have prompted active research in biological text mining to facilitate literature-based curation of molecular databases. Meanwhile, systems biology and bio-ontologies are emerging as critical tools in biological research where complex data in disparate resources are generated, integrated and analyzed. Both rely on literature for data annotation and analysis. The challenges facing us are to develop broadly utilized text mining tools and systems that need to involve both developers and users for system development and evaluation. iProLINK, extending from a previously developed text mining resource (Hu et al., 2004), is designed as a framework for linking text mining tools with ontology and systems biology. The framework focuses on text mining of protein-protein interaction, including the protein posttranslational modification such as phosphorylation, which can be applied to curation of molecular and ontological data and analysis of systems biology data.

The framework consists of two major components: a user interface for text mining of PPI from an integrated tool server and software modules to allow text mining outputs to be created, ranked, and used by the community. Use cases are presented for assessing the gaps and making recommendations for future development. The detailed components and case studies are described in a conference paper (Hu *et al*., 2008) [**Appendix IV**]. The iProLINK framework will benefit the Pathogen Mining project by not only maximally utilizing the different tools developed by the text mining community and providing an interface for community access, but also encouraging the use and application of these tools in the real-world applications such as assisting genomic and proteomic data analysis and pathogen data mining. We further organized a workshop during the PAG XVII (Plant and Animal Genome Conference) on "*Text Mining for Database Curation*" (Wu, 2009) (http://www.intl-pag.org/17/17-pir.html) [**Appendix V**] to present the iProLINK framework and to foster discussion on the development of text mining systems that address the needs of the biocuration and biological research community.

YEAR TWO

The objective of the second year was to use the Pathogen Mining System (**Figure 2**) to semi-automatically mine text for information on pathogenesis-related proteins, including host interacting proteins, and to use the text mining results in comprehensive functional analysis of high-throughput proteomic data from pathogenic and non-pathogenic *Burkholderia* strains grown

in different kinds of media. The scope of the project was defined under the Cooperative Research and Developm ent Agreem ent with USAMRIID entitled " *Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition*," which focused on prior 2DGE-MS (2D gel electr ophoresis-mass spectrometry) proteomics data from *Burkholderia* strains.

- ***Task1 (M13-15): Preliminary analysis of the Burkholderia proteomic space.*** Collecting data on *Burkholderia* strains and developing the scope of computational work performed to analyze *Burkholderia* proteomic data.
- ***Task2 (M16-18): Protein identification using MASCOT.*** Initial pro tein identif ication using MASCOT, and c onfirmation of the id entification through m anual checking and mapping of IDs back to 2-D gels.
- ***Task3 (M18-24): Annotation of identified proteins and integration data into iProXpress.*** Manual annotation of proteins identified using RACE-P interface, automated annotation of the *Burkholderia* genome using RAST, lite rature m ining of patho genic *Burkholderia* proteins, and using *i*ProXpress to perform mining and analysis of the data.

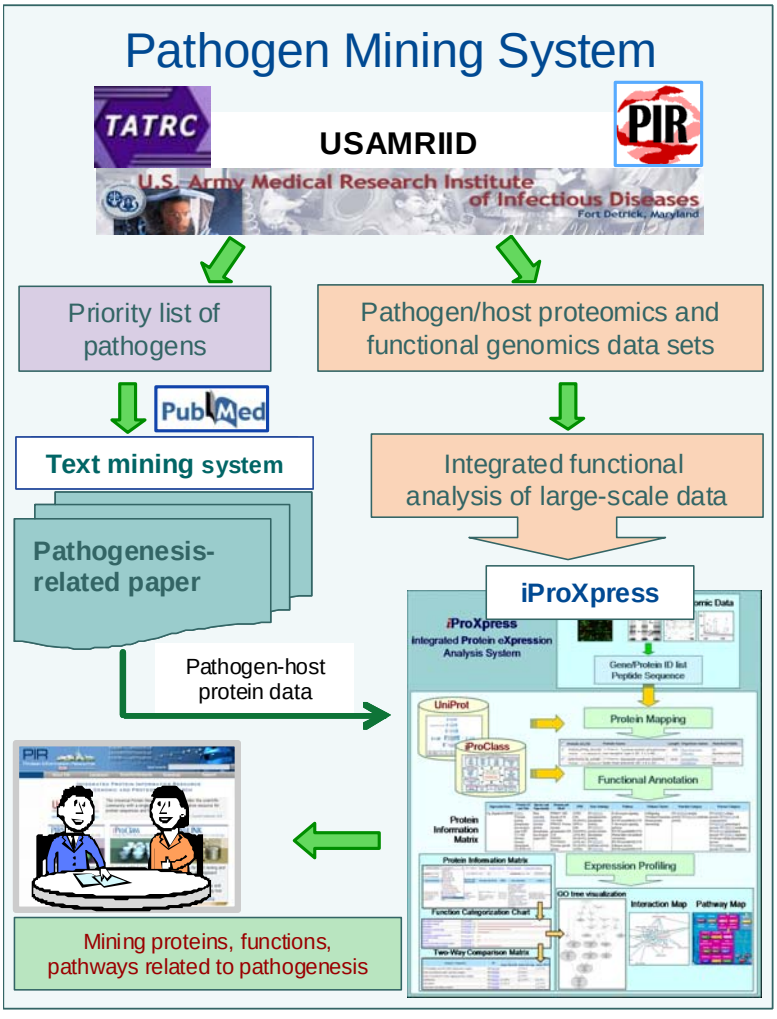


Figure 2. Pathogen Mining System.

I. Collaboration with USAMRIID research groups

The second year of the Pathogen Mining project was revised to focus on the collaborative work with the USAMRIID on bacterial pathogen proteomics data analysis using the iProXpress system developed at PIR (Huang et al, 2007). In the beginning of the project, we met with the USAMRIID research groups and discussed the research activities in their labs complimentary to this project. We agreed to analyze *Burkholderia* proteomic data obtained from growing strains *in vitro* and media mimicking *in vivo* conditions using literature mining and data mining methodologies that have been already developed at PIR. The hypothesis for the original proteomic experiments is as follows: Proteins important to Glanders disease pathology may be discovered through the comparative analysis of proteomes derived from bacteria cultured *in vitro* under conditions that partially simulate *in vivo* growth in the mammalian host.

To test this hypothesis *Burkholderia mallei* (human pathogen), *B. pseudomallei* (human pathogen), *B. vietnamiensis*(opportunistic human pathogen) and *B. thailandensis* (avirulent) were grown *in vitro* and simulated *in vivo* conditions (iron and calcium limited media) and their protein component was analyzed using 2-D gels. Proteins that were up-regulated and down-regulated were excised from the gels and was analyzed using MALDI-TOF-MS and peak lists were obtained. These peak lists were initially analyzed five years back against two *Burkholderia* proteomes. With the advent of twenty additional *Burkholderia* proteomes in the databases, we reanalyzed the peak list for better identification of the proteins. Then perform detailed analysis of the proteins. The objective of this collaboration was to use the integrated proteomics analysis system, iProXpress, coupled with the TATRC-funded project, Pathogen Mining System, to facilitate the re-evaluation and functional interpretation and hypothesis formulation from the legacy data.

1. Protein identification using MASCOT. The initial step towards protein identification is mapping the experimental organisms to specific *Burkholderia* strains in the NCBI taxonomy database (**Table 3**).

Table3. Mapping of strains used in experiment to NCBI taxonomy IDs
Each experiment was conducted with induced vs. uninduced growth condition of the organism.

Experiment/ Gel	Treatments (Uninduced/Induced)	Strains Used	Mapped to NCBI taxonomy ID
1	1/7	B. mallei GB8	B. mallei ATCC 23344 (taxid 243160)
2	2/8	B. mallei GB6	B. mallei (taxid 13373)
3	3/9	B. mallei GB5	B. mallei NCTC 10229 (taxid 412022)
4	4/10	B. pseudomallei 1126B	B. pseudomallei (taxid 320373)
5	5/11	B. thailandensis E254	B. thailandensis (taxid 271848)
6	6/12	B. vietnamensis FCO369	B. vietnamensis (taxid 269482)

We used Mascot Peptide Mass Fingerprint to identify proteins using the data files provided by Dr. Powell of USAMRIID. Dr. Powell, with close collaboration with us, confirmed the experimental conditions and the strains of *Burkholderia* (**Table 4**) for 200 MALDI spectra and they were used for Protein Identification with MASCOT search engine from the Proteomics Lab at the University of Delaware.

Table 4. List of *Burkholderia* Strains.

Organism Name	Taxonomy ID	# of files	# of Sequences
Burkholderia mallei (Pseudomonas mallei)	13373 77		4831
Burkholderia vietnamiensis (strain G4 / LMG 22486) (Burkholderia cepacia (strain R1808))	269482 35		7410
Burkholderia thailandensis (strain E264 / ATCC 700388 / DSM 13276 / CIP 106301)	271848 38		5563
Burkholderia pseudomallei (strain 668)	320373 26		7215
Burkholderia mallei (strain NCTC 10229)	412022 24		5309

Each spectrum is searched against its corresponding sequence databases using **Mascot Peptide Mass Fingerprint.** Search engine used is Mascot version 2.2, the information associated with Mascot Peptide Mass Fingerprint search is listed as follows:
Search Parameters:

Type of search: Peptide Mass Fingerprint
Enzyme: Trypsin
Fixed modifications: Carbamidomethyl (C)
Variable modifications: Oxidation (M)
Mass values: Monoisotopic
Protein Mass: Unrestricted
Peptide Mass Tolerance: ± 1.2 Da
Peptide Charge State: 1+
Max Missed Cleavages: 1

Protein score is -10*Log(P), where P is the probability that the observed match is a random event. For each spectra, we recorded all the identified proteins which are significant ($p<0.05$), not just top scoring protein. Overall, we reanalyzed the data using MASCOT, finished the final identification, functional annotation of the proteins and mapping them to the 2-D gel spots, and identified 173 unique UniProtKB IDs (W81XWH-07-2-0112_Supplement.xls) [**Appendix VI**].

2. Manual Annotation of Identified Proteins in RACE-P interface. Of the proteins identified, 31 are UniProtKB/Swiss-Prot entries and have already been manually curated. The next step was to annotate the other proteins. Rapid Annotation interfaCE for proteins (RACE-P) a web interface developed at PIR was used to annotate these proteins. The features of the RACE-P page are divided into following blocks:

BLOCK 1: Protein Name.
The name of the protein, short name, EC Number and synonyms are derived from publications and/or closely related UniProtKB/Swiss-Prot homolog. In twenty-one cases, no publications or closely related UniProtKB/Swiss-Prot homolog of the protein could be found; hence the name was derived from the author submitted data in the corresponding UniProtKB/TrEMBL entry.

BLOCK 2: Gene Name.
The Gene Name, synonym and Gene ID are derived from the publications, model organism database and/or the author submitted data in UniProtKB/TrEMBL. This block also contains a link to the gene record in Organism Database, www.burkholderia.com.

BLOCK 3: Bibliography.
This block contains information on the protein from Publications. Publications which describe experiments performed on the gene or protein are included in this section. Review articles are usually excluded.

BLOCK 4: Gene Ontology.
The Gene Ontology terms included in this section are derived only from publications.

BLOCK 5: Computational Analysis.
This is done to confirm and/or add new information. The tools used are European Molecular Biology Open Software Suite (EMBOSS) and PIR developed tools.

BLOCK 6: Protein family.
Families are created consisting of 50 BLAST hits with at least 70 % end-to-end overlap to the query and e-value better than 1.0E-10. The family names, synonym and EC Numbers are derived from the UniProtKB/SwissProt entries in the family. There were instances where no UniProtKB/SwissProt protein fit the family criteria.

A total of 66 proteins are annotated in the RACE-P, 24 of which have closely related UniProtKB/Swiss-Prot entries and 42 proteins do not. The remaining 76 are uncharacterized hypothetical proteins. An Excel file (W81XWH-07-2-0112_Supplement.xls) was created with the following column. **Table 5** shows the manual annotation of identified *Burkholderia* proteins, displaying partial sections of the annotated file.

Spectrum – The spectrum file name of the spot
Accession – UniProtKB ID mapped to the spot
Protein Name– Name given in the RACE-P annotation or UniProtKB/Swiss-Prot name
Mascot Protein – UniProtKB Accession, UniProt KB Protein Name and organism
Score, **Threshold**, **Expect**, **PeptideMatched** – Numerical results from mascot
Organism used in experiment – Organism name in Dr Powell's original file
Identified Strains – *Burkholderia* strains in the NCBI taxonomy database mapped to the TaxID
TaxID – Taxonomy ID
Un/Induced – Experimental condition; Induced/Uninduced
Sample – Gel Number
Spot Number – Spot on the gel
Comp. – Experimental comparison file name
Comp description – Experimental observation from comparisons between gels
OLD TIGR – Original TIGR annotation
Possible Associations searching by spot - Comments from Dr. Powell
Comments – Comments from Dr. Powel

3. Large-scale Annotation of the Five Burkholderia genomes using RAST. In addition the manual annotation of the proteins, we performed large-scale automated annotation using Rapid Annotation using Subsystem Technology (RAST) server. RAST is an automated service provided by one of PIR collaborators and it is useful in annotating bacterial and archaeal genomes (Aziz et al., 2008). We ran the 5 *Burkholderia* genome used in the 2-D gel experiments. The input into the web based program is the genome sequence in GenBank format. **Figure 3** represents the output for *Burkholderia mallei* ATCC 23344. The image provides an overview of

the genome and classifies the proteins into subsystems. The subsystem classification is different for the five genomes: for instance, *B. mallei* ATCC 23344 has 125 protein involved in virulence, while *B. vietnamiensis* has 204. The RAST result also provides additional annotation of the proteins identified.

Table 5. Manual annotation of identified *Burkholderia* proteins
Showing only partial sections of the annotated file (W81XWH-07-2-0112_Supplement.xls)

Spectrum	Accession	Protein Name	Identified Strains	Un/Induced	Sample	Spot Number
8.206_44_0001	Q62GL8	30S ribosomal protein S14	*B. mallei* ATCC 23344	Induced 8B		206
Usamrid__66_0001	A3NEG6	30S ribosomal protein S14	*B. pseudomallei* 668	Uninduced 4C		144
8.161_52_0001	Q62GL1	30S ribosomal protein S3	*B. mallei* ATCC 23344	Induced 8B		161
Usamrid__92_0001	A4JAP6	30S ribosomal protein S3	*B. vietnamiensis* G4	Uninduced 6B		
Usamrid__25_0001	Q62I82	60 kDa chaperonin	*B. mallei* ATCC 23344	Uninduced 1C		41
7.19_28_0001 Q62AV7		Cellulose synthase operon protein C	*B. mallei* ATCC 23344	Induced 7C		19
Usamrid__11_0001 Q62AV7		Cellulose synthase operon protein C	*B. mallei* ATCC 23344	Uninduced 1B		97
9.78_69_0001	A2SAM0	D-alanyl-D-alanine carboxypeptidase family protein	*B. mallei* NCTC 10229	Induced 9		78
Usamrid__57_0001	A2SAM0	D-alanyl-D-alanine carboxypeptidase family protein	*B. mallei* NCTC 10229	Uninduced 3		338
8.144_41_0001	Q62KW4	DNA helicase II	*B. mallei* ATCC 23344	Induced 8		144
6.118_05_0001 A4JV87		DNA topoisomerase	*B. vietnamiensis* G4	Uninduced 6C		118
12.154_47_0001	A4JAR6	DNA-directed RNA polymerase subunit alpha	*B. vietnamiensis* G4	Induced 12C		154
8.136_51_0001	Q62B16	Effector protein bopA	*B. mallei* ATCC 23344	Induced 8		136
Usamrid__43_0001	Q62B16	Effector protein bopA	*B. mallei* ATCC 23344	Uninduced 2C		102
Usamrid__77_0001	Q2T703	Effector protein bopA	*B. thailandensis* E264	Uninduced 5B		12
Usamrid__18_0001 Q62E W4		Ferredoxin--NADP reductase	*B. mallei* ATCC 23344	Uninduced 1B		166
9.3_67_0001 A2RZ L5		GMC oxidoreductase	*B. mallei* NCTC 10229	Induced 9B		3
Usamrid__94_0001	A4JPC6	GP32 family protein	*B. vietnamiensis* G4	Uninduced 6C		12
12.171_51_0001 A4JCC6		Guanosine-3,5-bis(diphosphate) 3-pyrophosphohydrolase	*B. vietnamiensis* G4	Induced 12C		171
Usamrid__55_0001	A2S4Y3	Isocitrate dehydrogenase [NADP]	*B. mallei* NCTC 10229	Uninduced 3C		324
11.208_29_0001 Q2SW A7		Putative aldehyde dehydrogenase	*B. thailandensis* E264	Induced 11C		208
Usamrid__71_0001	A3NED6	Putative PilN protein	*B. pseudomallei* 668	Uninduced 4C		221
7.27_30_0001	Q62K99	Putative Syringomycin biosynthesis enzyme	*B. mallei* ATCC 23344	Induced 7C		27
Usamrid__29_0001	Q62K99	Putative Syringomycin biosynthesis enzyme	*B. mallei* ATCC 23344	Uninduced 1C		111
Usamrid__85_0001	Q2T7B5	Putative Syringomycin biosynthesis enzyme	*B. thailandensis* E264	Uninduced 5B		90
6.69_01_0002 A4JNV7		Putative transposase	*B. vietnamiensis* G4	Uninduced 6		69
7.167b_17_0001	Q62IX8	Pyruvate dehydrogenase, E1 component	*B. mallei* ATCC 23344	Induced 7A		167
6.125_06_0001	A4JPQ7	Response regulator receiver protein	*B. vietnamiensis* G4	Uninduced 6C		125
Usamrid__84_0001 Q2T 916		Rhs1 protein	*B. thailandensis* E264	Uninduced 5B		75
Usamrid__55_0001 A2S2I 2		Ribonuclease R	*B. mallei* NCTC 10229	Uninduced 3C		324
Usamrid__65_0001 A3NAU5		Ribosome-recycling factor	*B. pseudomallei* 668	Uninduced 4C		134
Usamrid__08_0001 Q62JC7		Ribosome-recycling factor	*B. mallei* ATCC 23344	Uninduced 1B		52
12.83_42_0001	A4JT92	RNA-directed DNA polymerase (Reverse transcriptase)	*B. vietnamiensis* G4	Induced 12C		83
Usamrid__82_0001	Q2T1R7	Sensor histidine kinase	*B. thailandensis* E264	Uninduced 5B		63
7.179_39_0001 Q62CQ2		Sensor protein	*B. mallei* ATCC 23344	Induced 7C		179
7.153_23_0001 Q4V296		Sigma-54 dependent transcriptional regulator	*B. mallei* ATCC 23344	Induced 7B		153
11.112_22_0001	Q2SVC6	Succinyl-CoA:3-ketoacid-coenzyme A transferase subunit A	*B. thailandensis* E264	Induced 11C		112
10.27_85_0001 A3NAI 6		Trigger factor	*B. pseudomallei* 668	Induced 10B		27
9.162_63_0001 A2SBG2		Trigger factor	*B. mallei* NCTC 10229	Induced 9		162
Usamrid__09_0001 Q62JK6		Trigger factor	*B. mallei* ATCC 23344	Uninduced 1B		55

4. Literature mining for *Burkholdria* Pathogenicity. Less than 4000 articles were retrieved from PubMed using the keyword search "Burkholderia". We could not find publications on most of the thirty-two UniProtKB/TrEMBL proteins indentified in the previous quarter. Though we could not get much information based on our dataset, useful information on *Burkholderia* parthenogenesis can be derived from publications in PubMed. The literature mining tool developed in the first year of this project is used to search for scientific articles that show

Pathogen-Host Protein-Protein Interaction (Guixian *et al*., 2008). Of the 3712 articles searched, 34 research articles are recognized as positive. 16 were manually determined to be true positive for Pathogen-Host Protein-Protein Interaction. Most of the other articles describe pathogenic proteins and complexes without the host proteins they interact with.

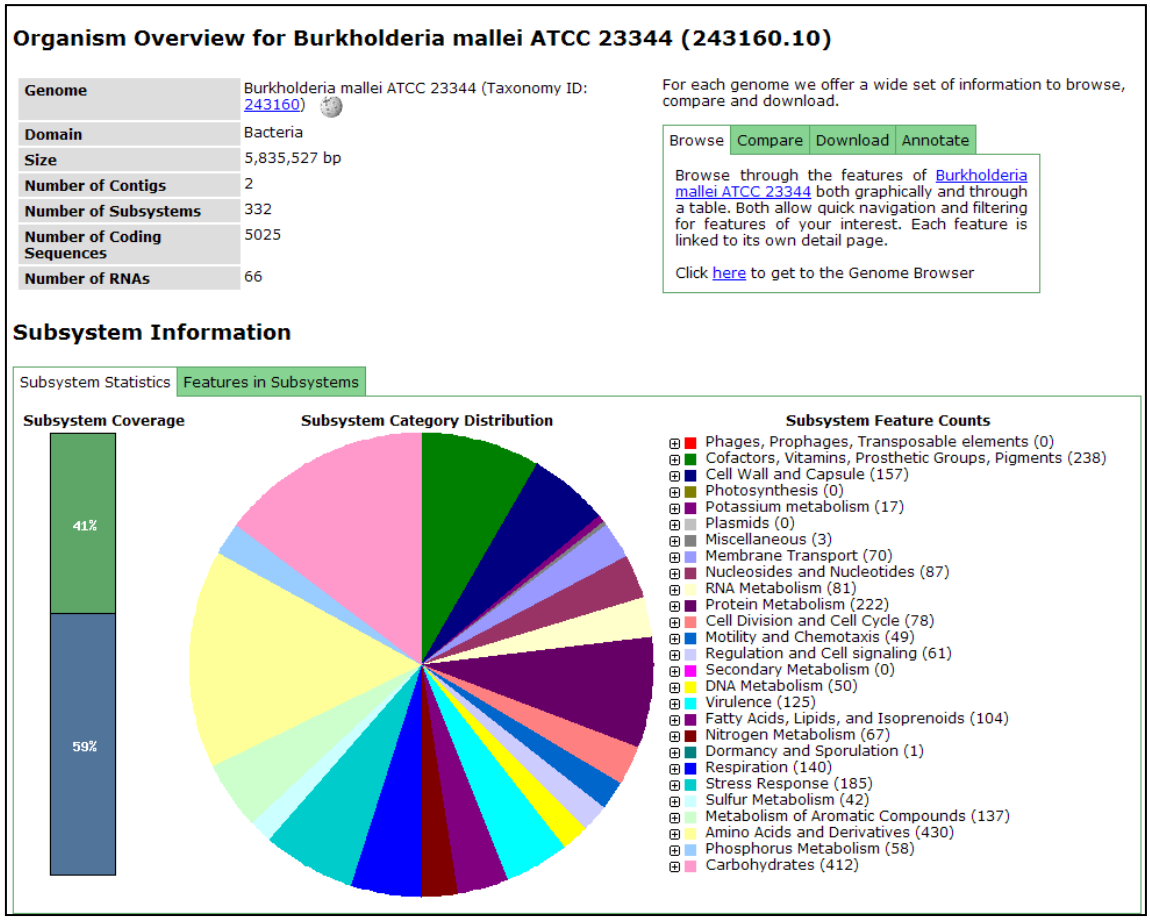
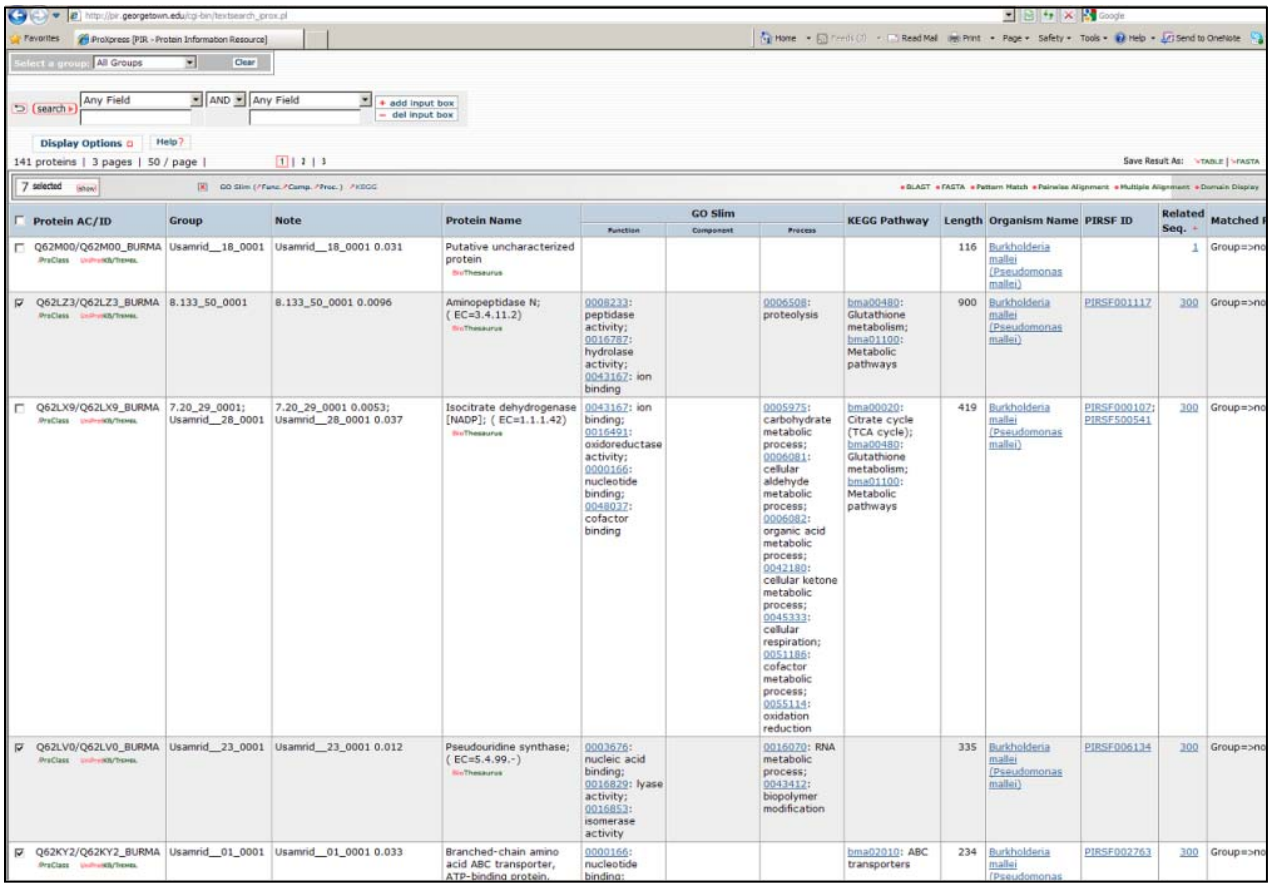


Figure 3. RAST results for *Burkholderia mallei* ATCC 23344.

5. Integration of proteins identified into iProXpress. *i*ProXpress (integrated Protein eXpression) is an integrated protein expression analysis system (Huang et al., 2007). 141 proteins identified by MASCOT with p-value <= 0.05 have been uploaded into *i*ProXpress (**Figure 4**). *i*ProXpress lists all entries and they are grouped by Spectrum. The data can be analyzed in the web interface. Functions of the system include Functional Profiling (sample in **Figure 5**), Protein Information Matrix and ID mapping. The information is accessible from http://pir.georgetown.edu/iproxpress/, under "Other data sets". The web pages are password protected and is provided to Dr. Powell for visualization of the annotated dataset.



Protein AC/ID	Group	Note	Protein Name	GO Slim (Function)	GO Slim (Component)	GO Slim (Process)	KEGG Pathway	Length	Organism Name	PIRSF ID	Related Seq.	Matched
Q62M00/Q62M00_BURMA	Usamrid__18_0001	Usamrid__18_0001 0.031	Putative uncharacterized protein					116	Burkholderia mallei (Pseudomonas mallei)		1	Group=>no
Q62LZ3/Q62LZ3_BURMA	8.133_50_0001	8.133_50_0001 0.0096	Aminopeptidase N; (EC=3.4.11.2)	0008233: peptidase activity; 0016787: hydrolase activity; 0043167: ion binding		0006508: proteolysis	bma00480: Glutathione metabolism; bma01100: Metabolic pathways	900	Burkholderia mallei (Pseudomonas mallei)	PIRSF001117	300	Group=>no
Q62LX9/Q62LX9_BURMA	7.20_29_0001; Usamrid__28_0001	7.20_29_0001 0.0053; Usamrid__28_0001 0.037	Isocitrate dehydrogenase [NADP]; (EC=1.1.1.42)	0043167: ion binding; 0016491: oxidoreductase activity; 0000166: nucleotide binding; 0048037: cofactor binding		0005975: carbohydrate metabolic process; 0006081: cellular aldehyde metabolic process; 0006082: organic acid metabolic process; 0042180: cellular ketone metabolic process; 0045333: cellular respiration; 0051186: cofactor metabolic process; 0055114: oxidation reduction	bma00020: Citrate cycle (TCA cycle); bma00480: Glutathione metabolism; bma01100: Metabolic pathways	419	Burkholderia mallei (Pseudomonas mallei)	PIRSF000107; PIRSF500541	300	Group=>no
Q62LV0/Q62LV0_BURMA	Usamrid__23_0001	Usamrid__23_0001 0.012	Pseudouridine synthase; (EC=5.4.99.-)	0003676: nucleic acid binding; 0016829: lyase activity; 0016853: isomerase activity		0016070: RNA metabolic process; 0043412: biopolymer modification		335	Burkholderia mallei (Pseudomonas mallei)	PIRSF006134	300	Group=>no
Q62KY2/Q62KY2_BURMA	Usamrid__01_0001	Usamrid__01_0001 0.033	Branched-chain amino acid ABC transporter, ATP-binding protein,	0000166: nucleotide binding;			bma02010: ABC transporters	234	Burkholderia mallei (Pseudomonas	PIRSF002763	300	Group=>no

Figure 4. *i*ProXpress protein information matrix (partially shown).

GO:0000166	nucleotide binding	15
GO:0001882	nucleoside binding	12
GO:0003676	nucleic acid binding	27
GO:0003824	catalytic activity	9
GO:0004386	helicase activity	2
GO:0004803	transposase activity	1
GO:0004871	signal transducer activity	1
GO:0005198	structural molecule activity	5
GO:0005215	transporter activity	1
GO:0005488	binding	4
GO:0005515	protein binding	4
GO:0008233	peptidase activity	2
GO:0009055	electron carrier activity	3
GO:0016491	oxidoreductase activity	9
GO:0016740	transferase activity	9
GO:0016787	hydrolase activity	12
GO:0016829	lyase activity	1
GO:0016853	isomerase activity	5
GO:0016874	ligase activity	4
GO:0019842	vitamin binding	1
GO:0030528	transcription regulator activity	3
GO:0031406	carboxylic acid binding	1
GO:0043167	ion binding	7
GO:0045182	translation regulator activity	3
GO:0048037	cofactor binding	4
GO:0051540	metal cluster binding	2
	All	59

Figure 5. Functional profile of 141 proteins, showing GO Slim functional categories.

KEY RESEARCH ACCOMPLISHMENTS:

- We manually curated pathogen-host PPI literature data sets that are necessary for the machine learning method as well as beneficial to the text mining community when becoming publicly available.
- We developed and evaluated the SVM methods for classifying the abstracts with PH-PPI information, whose overall performance is best when using sentence level training and feature selection.
- We identified and evaluated existing public text mining tools such as PIE that can be augmenting the Pathogen Mining System.
- We initiated a community collaborative effort under the iProLINK framework, which will be of great benefit to the Pathogen Mining System.
- We established close collaborations with USAMRIID research groups to analyze pathogen genomic and proteomic data that will take advantage of the PH-PPI text mining.
- We identified proteins from prior 2DGE-MS *Burkholderia* proteomics data.
- We manually annotated the proteins in the RACE-P interface of iProClass.
- We integrated the identified proteins into iProXpress to perform mining and analysis of the data.

REPORTABLE OUTCOMES:

1. Cooperative Research and Development Agreement between Georgetown University and USAMRIID
2. Three research papers were generated from the project, reporting the SVM-based PH-PPI text mining system (Xu et al., 2008; Yin et al., 2009) and an integrated text mining framework for text mining and biology communities (Hu et al., 2008).
3. A workshop presentation at the 2009 PAG XVII (Plant and Animal Genome Conference) on the iProLINK framework (Wu, 2009) (http://www.intl-pag.org/17/17-pir.html).

CONCLUSIONS:

Biomedical literature represents the primary source of experimental data, and developing text mining systems for mining such data for pathogens of biodefense relevance is the main objective for the first year of the project. We focus on text mining of the host-pathogen protein-protein interactions. We developed an SVM-based automated system to identify MEDLINE abstracts containing HP-PPI information. We observed that feature selection was effective not only in reducing the dimensionality of features to build a compact system, but also in improving document classification performance. We also observed abstract-level systems and sentence-level systems yielded different classification of MEDLINE abstracts, and the combination of these systems could improve the overall document classification. To augment the SVM-based PH-PPI mining methods, we also explored the public text mining tools for the PH-PPI mining. We performed preliminary evaluation on the PPI extraction tool PIE, and the results showed encouraging performance at least at the abstract level, suggesting that PIE can be potentially integrated into the Pathogen Mining System for improving the overall text mining capabilities of the system. Exploring public text mining tools is also part of the initiative by PIR in order to develop a basic framework to bring together the text mining and biological communities to better

develop text mining tools for real-world applications. Our second year tasks focused on the identification, annotation and analysis of and proteomic data for pathogens of biodefense and military relevance.

REFERENCES:

1. Fernández JM, Hoffmann R, Valencia A. (2007) iHOP web services. *Nucleic Acids Res.* 35 (Web Server issue): W21-26.
2. Hirschman L, Yeh A, Blaschke C, Valencia A. (2005) Overview of BioCreative: Critical Assessment of Information Extraction for Biology. *BMC Bioinformatics* 6 (Suppl 1): S1.
3. Hu ZZ, Cohen KB, Hirschman L, Valencia A, Liu H, Giglio MG, Wu C H. (2008) iProLINK: A Framework for Linking Text Mining with Ontology and System s Biology. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine* (BIBM 2008), pp 467-472.
4. Wu CH, Hirschman L. Linking Text Mining with Ontology and System s Biology for Database Curation. (Workshop abstract) Plant and Animal Genome Conference (PAG) XVII, San Diego, CA, January 10-14, 2009
5. Hu ZZ, Mani I, Hermoso V, Liu H, Wu CH. (2004) iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem* 28: 409-416.
6. Huang H, Hu ZZ, Arighi CN, Wu CH. (2007). Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Front Biosci.* 12: 5071-5088.
7. Kim S, Shin SY, Lee IH, Kim SJ, Sriram R, Zhang BT. (2008) PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.* 36 (Web Server issue): W411-415.
8. Leitner F, Krallinger M, Rodriguez-Penagos C, Hakenberg J, Plake C, *et al.* (2008) Introducing meta-services for biomedical information extraction. *Genome Biology* 9 (Suppl 2): S6.
9. Wilbur J, Smith L, Tanabe L. (2007) BioCreative 2. Gene Mention Task, In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pp. 7–16.
10. Xu G, Yin L, Torii M, Niu Z, Wu CH, Hu Z, Liu H. (2008). Document classification for mining host pathogen protein- protein interactions. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine* (BIBM 2008), pp 461-466.
11. Yin L, Xu G, Torii M, Niu Z, Wu CH, Hu Z, Liu H. (2009) Document classification for mining host pathogen protein-protein interactions. *Artificial Intelligence in Medicine* (in press)
12. Aziz RK, Bartels D, Best AA, DeJongh M, Disz T, Edwards RA, Formsma K, Gerdes S, Glass EM, Kubal M, Meyer F, Olsen GJ, Olson R, Osterman AL, Overbeek RA, McNeil LK, Paarmann D, Paczian T, Parrello B, Pusch GD, Reich C, Stevens R, Vassieva O, Vonstein V, Wilke A, Zagnitko O. (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics* 9: 75.

APPENDICES:

I. Cooperative Research And Developm ent Agreem ent (USAMRMC Control No: W81XWH-09-0003) between Georgetown Un iversity and USAM RIID (US Ar my Medical Research Inst itute of Infectious Dis ease), entitled " *Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition*"

II. Xu G, Yin L, Torii M, Niu Z, W u CH, Hu Z, Liu H. (2008). Docum ent classification for mining host pathogen protein- protein interactions. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine* (BIBM 2008), pp 461-466.
[http://www.com puter.org/portal/web/csdl/doi/10.1109/BIBM.2008.66]

III. Yin L, Xu G, Torii M, Niu Z, Wu CH, Hu Z, Liu H. (2009) Docum ent classification for mining host pathogen protein-protein interactions. *Artificial Intelligence in Medicine* (in press)

IV. Hu ZZ, Cohen KB, Hi rschman L, Valenc ia A, Liu H, Giglio MG, W u C H. (2008) iProLINK: A Fra mework for Linking Text Mining with Ontology and System s Biology. In *Proceedings of IEEE International Conference on Bioinformatics and Biomedicine* (BIBM 2008), pp 467-472.
[http://www.computer.org/portal/web/csdl/doi/10.1109/BIBM.2008.73]

V. PAG XVII (Plant and A nimal Genome Conference): PIR (Protein Information Resource) Workshop on Text Mining for Database Curation (Wu, 2009)
[http://www.intl-pag.org/17/17-pir.html] and [http://www.intl-pag.org/17/abstracts/]

VI. W81XWH-07-2-0112_Supplement.xls: Protei n identification and m anual annotation results of *Burkholderia* proteins reanalyzed from USAMRIID proteomic data

COVER SHEET
COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT

[NOTE: This Cover Sheet is for internal management purposes only. It is not part of the Agreement and neither party is bound to anything contained in it]

Title: Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition

Effective Date : 10-07-2008
Expiration Date: 10-07-2010

USAMRMC Control No. W81XWH-09-0003
DA/TTPO Control No.

Primary NTIS Subject Code/Title: 57K
Secondary NTIS Subject Code/Title:
STO Code/Title:
Oracle #: 92951
Concurrence obtained from appropriate RAD/USAMMDA/CBMS-JPMO program managers: YES/RAD 4

Laboratory:
USAMRIID
MCMR-UIZ-D
1425 Porter Street
Fort Detrick, MD 21702-5011
Voice Phone: 301-619-6886 FAX Phone: 301-619-8379

Lab's Technical POC
Dr. Bradford Powell
USAMRIID / MCMR-UIB
1425 Porter Street, Fort Detrick, MD 21702-5011
Voice Phone: 301-619-4933 FAX Phone: 301-619-2152
Email: Bradford.powell@amedd.army.mil

Lab's Legal Counsel:
Commander, U.S. Army Medical Research and Materiel Command
ATTN: MCMR-JA (Mr. Robert L. Charles)
Fort Detrick, Frederick, MD 21701-5012
Voice Phone: 301-619-2065 FAX Phone: 301-619-5034

Cooperator's POCs:
Dr. Cathy H. Wu (Scientific POC)
Georgetown University Medical Center
3300 Whitehaven Street, NW; Suite 1200
Washington, DC 20007
Phone: 202-687-1039 Fax: 202-687-0057
Email: wuc@georgetown.edu

Silvana T. Alcocer (Administrative)
IP & Contract Administrator; Office of Technology Commercialization
Georgetown University
3300 Whitehaven Street, NW; Suite 1500
Washington, DC 20007
Phone: 202-687-0843 Fax: 202-687-3111
Email: alcocers@georgetown.edu

Summary: In this collaboration, USAMRIID will provide MS data comprising protein lists and other information of relevance for matched data sets to be re-analyzed by the Pathogen Mining system.

Bob Charles reviewed 9/15/2008.

A COOPERATIVE RESEARCH AND DEVELOPMENT AGREEMENT

Between

Georgetown University
37th and O Streets, NW
Washington, District of Columbia 20057
(Cooperator)
and
U.S. Army Medical Research Institute of Infectious Diseases
Fort Detrick, Maryland 21702-5011
(Laboratory)

Article 1. Background

1.00 This Agreement is entered into under the authority of the Federal Technology Transfer Act of 1986, 15 U.S.C. 3710a, et seq., between the Cooperator and the Laboratory, the parties to this Agreement.

1.01 Laboratory, on behalf of the U.S. Government, and Cooperator desire to cooperate in research and development on Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition according to the attached Statement of Work (SOW) described in Appendix A. NOW, THEREFORE, the parties agree as follows:

Article 2. Definitions

2.00 The following terms are defined for this Agreement as follows:

2.01 "Agreement" means this cooperative research and development agreement.

2.02 "Invention" and "Made" have the meanings set forth in Title 15 U.S.C. Section 3703(9) and (10).

2.03 "Proprietary Information" means information marked with a proprietary legend which embodies trade secrets developed at private expense or which is confidential business or financial information, provided that such information:

(i) is not generally known, or which becomes generally known or available during the period of this Agreement from other sources without obligations concerning their confidentiality;

(ii) has not been made available by the owners to others without obligation concerning its confidentiality; and

(iii) is not already available to the receiving party without obligation concerning its confidentiality.

(iv) is not independently developed by or on behalf of the receiving party, without reliance on the information received hereunder.

2.04 "Subject Data" means all recorded information first produced in the performance of this Agreement.

2.05 "Subject Invention" means any Invention Made as a consequence of, or in relation to, the performance of work under this Agreement.

Article 3. Research Scope and Administration

3.00 Statement of Work. Research performed under this Agreement shall be performed in accordance with the SOW incorporated as a part of this Agreement at Appendix A. It is agreed that any descriptions, statements, or specifications in the SOW shall be interpreted as goals and objectives of the services to be provided under this Agreement and not requirements or warranties. Laboratory and Cooperator will endeavor to achieve the goals and objectives of such services; however, each party acknowledges that such goals and objectives, or any anticipated schedule of performance, may not be achieved.

3.01 Review of Work. Periodic conferences shall be held between the parties for the purpose of reviewing the progress of work. It is understood that the nature of this research is such that completion within the period of performance specified, or within the limits of financial support allocated, cannot be guaranteed. Accordingly, all research will be performed in good faith.

3.02 Principal Investigator. Any work required by the Laboratory under the SOW will be performed under the supervision of Dr. Bradford Powell, U.S. Army Medical Research Institute of Infectious Diseases (USAMRIID) 1425 Porter Street, Fort Detrick, MD 21702-5011, Phone: 301-619-4933, Fax 301-619-2152 and Email: bradford.powell@amedd.army.mil, who, as co-principal investigator has responsibility for the scientific and technical conduct of this project on behalf of the Laboratory. Any work required by the Cooperator under the SOW will be performed under the supervision of Dr. Cathy H. Wu, Georgetown University Medical Center, 3300 Whitehaven Street, NW, Suite 1200, Washington, DC 20007, Phone: 202-687-1039, Fax: 202-687-0057, and Email: wuc@georgetown.edu, who, as co-principal investigator has responsibility for the scientific and technical conduct of this project on behalf of the Cooperator.

3.03 Collaboration Changes. If at any time the co-principal investigators determine that the research data dictates a substantial change in the direction of

the work, the parties shall make a good faith effort to agree on any necessary change to the SOW and make the change by written notice to the addresses listed in section 12.05 Notices.

3.04 Final Report. The parties shall prepare a final report of the results of this project within six months after completing the SOW.

Article 4. Ownership and Use of Physical Property

4.01 Ownership of Materials or Equipment. All materials or equipment developed or acquired under this Agreement by the parties shall be the property of the party which developed or acquired the property, except that government equipment provided by Laboratory (1) which through mixed funding or mixed development must be integrated into a larger system, or (2) which through normal use at the termination of the Agreement has a salvage value that is less than the return shipping costs, shall become the property of Cooperator.

4.02 Use of Provided Materials. Both parties agree that any materials relating to them which were provided by one party to the other party will be used for research purposes only. The materials shall not be sold, offered for sale, used for commercial purposes, or be furnished to any other party without advance written approval from the Provider's official signing this Agreement or from another official to whom the authority has been delegated, and any use or furnishing of material shall be subject to the restrictions and obligations imposed by this Agreement.

Article 5. Patent Rights

5.00 Reporting. The parties shall promptly report to each other all Subject Inventions reported to either party by its employees. All Subject Inventions Made during the performance of this Agreement shall be listed in the Final Report required by this Agreement.

5.01 Cooperator Employee Inventions. Laboratory waives any ownership rights the U.S. Government may have in Subject Inventions Made by Cooperator employees and agrees that Cooperator shall have the option to retain title in Subject Inventions Made by Cooperator employees. Cooperator shall notify Laboratory promptly upon making this election and agrees to timely file patent applications on Cooperator's Subject Invention at its own expense. Cooperator agrees to grant to the U.S. Government on Cooperator's Subject Inventions a nonexclusive, nontransferable, irrevocable, paid-up license in the patents covering a Subject Invention, to practice or have practiced, throughout the world by, or on behalf of the U.S. Government. The nonexclusive license shall be evidenced by a confirmatory license agreement prepared by Cooperator in a form satisfactory to Laboratory.

5.02 Laboratory Employee Inventions. Laboratory shall have the initial option to retain title to, and file patent application on, each Subject Invention Made by its employees. The Laboratory agrees to grant an exclusive license to any invention arising under this Agreement to which it has ownership to the Cooperator in accordance with Title 15 U.S. Code Section 3710a, on terms negotiated in good faith. Any invention arising under this Agreement is subject to the retention by the U.S. Government of nonexclusive, nontransferable, irrevocable, paid-up license to practice, or have practiced, the invention throughout the world by or on behalf of the U.S. Government.

5.03 Joint Inventions. Any Subject Invention patentable under U.S. patent law which is Made jointly by Laboratory employees and Cooperator employees under the Scope of Work of this Agreement shall be jointly owned by the parties. The parties shall discuss together a filing strategy and filing expenses related to the filing of the patent covering the Subject Invention. If a party decides not to retain its ownership rights to a jointly owned Subject Invention, it shall offer to assign such rights to the other party, pursuant to Paragraph 5.05, below.

5.04 Government Contractor Inventions. In accordance with 37 Code of Federal Regulations 401.14, if one of Laboratory's Contractors conceives an invention while performing services at Laboratory to fulfill Laboratory's obligations under this Agreement, Laboratory may require the Contractor to negotiate a separate agreement with Cooperator regarding allocation of rights to any Subject Invention the Contractor makes, solely or jointly, under this Agreement. The separate agreement (i.e., between the Cooperator and the Contractor) shall be negotiated prior to the Contractor undertaking work under this Agreement or, with the Laboratory's permission, upon the identification of a Subject Invention. In the absence of such a separate agreement, the Contractor agrees to grant the Cooperator an option for a license in Contractor's inventions of the same scope and terms set forth in this Agreement for inventions made by Laboratory employees.

5.05 Filing of Patent Applications. The party having the right to retain title to, and file patent applications on, a specific Subject Invention may elect not to file patent applications, provided it so advises the other party within 90 days from the date it reports the Subject Invention to the other party. Thereafter, the other party may elect to file patent applications on the Subject Invention and the party initially reporting the Subject Invention agrees to assign its ownership interest in the Subject Invention to the other party.

5.06 Patent Expenses. The expenses attendant to the filing of patent applications shall be borne by the party filing the patent application. Each party shall provide the other party with copies of the patent applications it files on any Subject Invention, along with the power to inspect and make copies of all documents retained in the official patent application files by the applicable patent

office. The parties agree to reasonably cooperate with each other in the preparation and filing of patent applications resulting from this Agreement.

Article 6. Exclusive License

6.00 Grant. The Laboratory agrees to grant to the Cooperator an exclusive license in each U.S. patent application, and patents issued thereon, covering a Subject Invention, which is filed by the Laboratory subject to the reservation of a nonexclusive, nontransferable, irrevocable, paid-up license to practice and have practiced the Subject Invention on behalf of the United States.

6.01 Exclusive License Terms. The Cooperator shall elect or decline to exercise its right to acquire an exclusive license to any Subject Invention within six months of being informed by the Laboratory of the Subject Invention. The specific royalty rate and other terms of license shall be negotiated promptly in good faith and in conformance with the laws of the United States.

Article 7. Background Patent(s)

7.00 Laboratory Background Patent(s): Laboratory has filed patent application(s), or is the assignee of issued patent(s) which contain(s) claims that are related to research contemplated under this Agreement. No license(s) to this/these patent applications or issue patents is/are granted under this Agreement, and this/these application(s) and any continuations to it/them are specifically excluded from the definitions of "Subject Invention" contained in this Agreement.

7.01 Cooperator Background Patent(s): Cooperator has filed patent application(s), or is the assignee of issued patent(s) which contain(s) claims that are related to research contemplated under this Agreement. No license(s) to this/these patent applications or issue patents is/are granted under this Agreement, and this/these application(s) and any continuations to it/them are specifically excluded from the definitions of "Subject Invention" contained in this Agreement.

Article 8. Subject Data and Proprietary Information

8.00 Subject Data Ownership. Subject Data shall be jointly owned by the parties. Each party, upon request to the other party, shall have the right to review and to request delivery of all Subject Data, and delivery shall be made to the requesting party within two weeks of the request, except to the extent that such Subject Data are subject to a claim of confidentiality or privilege by a third party.

8.01 Proprietary Information/Confidential Information. Each party shall place a proprietary notice on all information it delivers to the other party under

this Agreement that it asserts is proprietary. The parties agree that any Proprietary Information or Confidential Information furnished by one party to the other party under this Agreement, or in contemplation of this Agreement, shall be used, reproduced and disclosed by the receiving party only for the purpose of carrying out this Agreement, and shall not be released by the receiving party to third parties unless consent to such release is obtained from the providing party.

8.02 Army limited-access database. Notwithstanding anything to the contrary in this Article, the existence of established CRADAs specifying areas of research and their total dollar amounts may be documented on limited access, password-protected websites of the U.S. Army Medical Research and Materiel Command (the parent organization of Laboratory), to provide the Command's leadership with a complete picture of military research efforts.

8.03 Laboratory Contractors. Cooperator acknowledges and agrees to allow Laboratory's disclosure of Cooperator's proprietary information to Laboratory's Contractors for the purposes of carrying out this Agreement. Laboratory agrees that it has or will ensure that its Contractors are under written obligation not to disclose Cooperator's proprietary information, except as required by law or court order, before Contractor employees have access to Cooperator's proprietary information under this Agreement.

8.04 Release Restrictions. Laboratory shall have the right to use all Subject Data for any Governmental purpose, but shall not release Subject Data publicly except: (i) Laboratory in reporting on the results of research may publish Subject Data in technical articles and other documents to the extent it determines to be appropriate; and (ii) Laboratory may release Subject Data where release is required by law or court order. The parties agree to confer prior to the publication of Subject Data to assure that no Proprietary Information is released and that patent rights are not jeopardized. Prior to submitting a manuscript for review which contains the results of the research under this Agreement, or prior to publication if no such review is made, each party shall be offered an ample opportunity to review any proposed manuscript and to file patent applications in a timely manner.

8.05 FDA Documents. If this Agreement involves a product regulated by the U.S. Food and Drug Administration (FDA), then the Cooperator or the U.S. Army Medical Research and Materiel Command, as appropriate, may file any required documentation with the FDA. In addition, the parties authorize and consent to allow each other or their contractors or agents access to, or to cross-reference, any documents filed with the FDA related to the product.

Article 9. Termination

9.00 Termination by Mutual Consent. Cooperator and Laboratory may elect to terminate this Agreement, or portions thereof, at any time by mutual consent.

9.01 Termination by Unilateral Action. Either party may unilaterally terminate this entire Agreement at any time by giving the other party written notice, not less than 30 days prior to the desired termination date.

9.02 Termination Procedures. In the event of termination, the parties shall specify the disposition of all property, patents and other results of work accomplished or in progress, arising from or performed under this Agreement by written notice. Upon receipt of a written termination notice, the parties shall not make any new commitments and shall, to the extent feasible, cancel all outstanding commitments that relate to this Agreement. Notwithstanding any other provision of this Agreement, any exclusive license entered into by the parties relating to this Agreement shall be simultaneously terminated unless the parties agree to retain such exclusive license.

Article 10. Disputes

10.00 Settlement. Any dispute arising under this Agreement which is not disposed of by agreement of the principal investigators shall be submitted jointly to the signatories of this Agreement. A joint decision of the signatories or their designees shall be the disposition of such dispute. However, nothing in this section shall prevent any party from pursuing any and all administrative and/or judicial remedies which may be allowable.

Article 11. Liability

11.00 Property. Neither party shall be responsible for damages to any property provided to, or acquired by, the other party pursuant to this Agreement.

11.01 No Warranty. The parties make no express or implied warranty as to any matter whatsoever, including the conditions of the research or any Invention or product, whether tangible or intangible, Made, or developed under this agreement, or the ownership, merchantability, or fitness for a particular purpose of the research or any Invention or product. The parties further make no warranty that the use of any invention or other intellectual property or product contributed, made or developed under this Agreement will not infringe any other United States or foreign patent or other intellectual property right. In no event will any party be liable to any other party for compensatory, punitive, exemplary or consequential damages.

Article 12. Miscellaneous

12.00 Governing Law. This Agreement shall be governed by the laws of the United States Government.

12.01 Export Control and Biological Select Agents and Toxins. The obligations of the parties to transfer technology to one or more other parties, provide technical information and reports to one or more other parties, and otherwise perform under this Agreement are contingent upon compliance with applicable United States export control laws and regulations. The transfer of certain technical data and commodities may require a license from a cognizant agency of the United States Government or written assurances by the Parties that the Parties shall not export technical data, computer software, or certain commodities to specified foreign countries without prior approval of an appropriate agency of the United States Government. The Parties do not, alone or collectively, represent that a license shall not be required, nor that, if required, it shall be issued. In addition, where applicable, the parties agree to fully comply with all laws, regulations, and guidelines governing biological select agents and toxins.

12.02 Independent Contractors. The relationship of the parties to this Agreement is that of independent contractors and not as agents of each other or as joint venturers or partners.

12.03 Use of Name or Endorsements. (a) The parties shall not use the name of the other party on any product or service which is directly or indirectly related to either this Agreement or any patent license or assignment agreement which implements this Agreement without the prior approval of the other party. (b) By entering into this Agreement, Laboratory does not directly or indirectly endorse any product or service provided, or to be provided, by Cooperator, its successors, assignees, or licensees. Cooperator shall not in any way imply that this Agreement is an endorsement of any such product or service. Press releases or other public releases of information shall be coordinated between the parties prior to release, except that the Laboratory may release the name of the Cooperator and the title of the research without prior approval from the Cooperator.

12.04 Survival of Specified Provisions. The rights specified in provisions of this Agreement covering Patent Rights, Subject Data and Proprietary Information, and Liability shall survive the termination or expiration of this Agreement.

12.05 Notices. All notices pertaining to or required by this Agreement shall be in writing and shall be signed by an authorized representative addressed as follows:

If to Cooperator: Georgetown University
Office of Technology Commercialization
3300 Whitehaven Street, N.W.
Harris Building, Suite 1500
Washington, DC 20007
Phone: 202-687-2702
Fax: 202-687-3111 (if by Fed Ex or courier)

Or use

Office of Technology Commercialization
Georgetown University
Box 571408
Washington, DC 20057-1408 (for US Mail)

If to Laboratory: USAMRIID
Business Plans and Programs Office
1425 Porter Street
Fort Detrick, MD 21702-5011
Phone: 301-619-6886 Fax: 301-619-8379

Any party may change such address by notice given to the other in the manner set forth above.

Article 13. Duration of Agreement and Effective Date

13.01 Effective Date. This Agreement shall enter into force as of the date it is signed by the last authorized representative of the parties.

13.02 Signature Execution. This Agreement may be executed in one or more counterparts by the parties by signature of a person having authority to bind the party, which may be by facsimile signature, each of which when executed and delivered, by facsimile transmission, mail, or email delivery, will be an original and all of which will constitute but one and the same Agreement.

13.03 Expiration Date. This Agreement will automatically expire two (2) years from effective date unless it is revised by written notice and mutual agreement.

IN WITNESS WHEREOF, the Parties have caused this agreement to be executed by their duly authorized representatives as follows:

For the Cooperator: Georgetown University
(Organization)

(Signature)

Claudia Cherney Stewart, Ph.D.

DATE 9/17/08 Vice President, Office of Technology Commercialization

For the U.S. Government: U. S. Army Medical Research Institute of Infectious Diseases

(Signature)

John P. Skvorak
Colonel, Veterinary Corps
Commanding

DATE 7 OCT 08

For the USAMRIID Principal Investigator:

I hereby acknowledge the terms and conditions of this Agreement:

DATE 24 Sept '08

(Signature)

Dr. Bradford Powell
(Printed Name)

(CRADA) APPENDIX A
STATEMENT OF WORK

Title: "Reanalysis and Functional Interpretation of Proteomics Data from Bacterial Cells under Simulated Growth Condition"

Background/Objectives:

Prior bacterial proteomics data needs to be reanalyzed due to the updates to the relevant bacterial protein databases and/or annotations, as well as accumulation of literature information regarding prior unknown genes. The objective of this collaboration is to use the integrated proteomics analysis system, iProXpress developed at PIR, coupled with the current TATRC- funded project, Pathogen Mining System, to facilitate the re-evaluation and functional interpretation and hypothesis formulation from the legacy proteomics data.

Prior 2DGE-MS proteomics data from Burkholderia strains grown under simulated host growth condition will be reanalyzed using iProXpress system for: 1) up-to-date functional assignment of bacterial protein annotations; 2) annotations of homologous proteins from other related pathogens of interests; 3) function and pathway analysis of the bacteria under given growth conditions.

Collaboration:

Laboratory agrees to:

- Provide MS data comprising protein lists and other information of relevance for matched data sets to be re-analyzed by the Pathogen Mining system.

Cooperator agrees to:

- Integrate all available annotations for proteins of Burkholderia and related bacteria into the iProXpress system, including biological pathways and experimental protein-protein interactions.
- Integrate into iProXpress the text mining results on pathogenesis proteins from the Pathogen Mining System.
- Incorporate the experimental Burkholderia proteomics data into the iProXpress system, and perform function and pathway analysis of the data.
- Enhance the iProXpress analysis interface based on the specific needs.

Document Classification for Mining Host Pathogen Protein-Protein Interactions

Guixian Xu[1,2,3,*], Lanlan Yin[1,*], Manabu Torii[4], Zhendong Niu[2], Cathy Wu[5], Zhangzhi Hu[5] and Hongfang Liu[1]

1. DBBB, Georgetown University Medical Center, Washington DC, USA
2. College of Computer Science, Beijing Institute of Technology, Beijing, China
3. College of Information Engineering, Central University for Nationalities, Beijing, China
4. ISIS Center, Georgetown University Medical Center, Washington DC, USA
5. PIR, Georgetown University Medical Center, Washington, DC, USA
{gx6,ly46,mt352,zh9,wuc,hl224}@georgetown.edu; zniu@bit.edu.cn

Abstract

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, a flood of molecular data about human pathogens has been generated and maintained in disparate databases. However, scientific findings regarding these pathogens and their host responses are buried in the growing volume of biomedical literature and there is an urgent need to mine information pertaining to pathogenesis-related proteins especially host-pathogen protein-protein interactions from literature. In this paper, we report our exploration of developing an automated system to identify MEDLINE abstracts referring to host-pathogen protein-protein interactions. An annotated corpus consisting of 1,360 MEDLINE abstracts was generated. With this corpus, we developed and evaluated document classification systems using support vector machines (SVMs). We also investigated the effects of feature selection using the information gain (IG) measure. Document classification systems were designed at two levels, abstract-level and sentence-level. We observed that feature selection was effective not only in reducing the dimensionality of features to build a compact system, but also in improving document classification performance. We also observed abstract-level systems and sentence-level systems yielded different classification of MEDLINE abstracts, and the combination of these systems could improve the overall document classification.

1. Introduction

Due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there have been major initiatives for large-scale genomic and proteomic projects to study the basic biology and disease-causing mechanisms of human pathogens [1, 2]. As a result, a flood of molecular data is being generated, but important scientific discoveries regarding these pathogens and their host responses are often buried under the increasing volume of biomedical literature.

Over the years, biomedical literature mining advanced greatly. In this paper, our investigation focused on the development of an automated system to identify research articles describing pathogenicity and host-pathogen protein-protein interactions. Our goal is to facilitate literature-based curation of pathogenesis-related proteins in UniProt Knowledgebase (UniProtKB) [3] by incorporating pathogenesis information extracted from literature and promoting basic understanding of virulence and pathogenicity factors as well as host-interacting proteins of human pathogens. Such knowledge will facilitate the development of preventative and therapeutic strategies against human pathogens.

In the following, we first describe the research background and related work. The experimental method is introduced next. We then present the results and discussion, and conclude our work.

2. Background and related work

The task considered in this study is a special

* Equal contribution to the work.

case of identifying papers that describe protein-protein interactions (PPIs). There are several components in developing an automated literature mining system, including the construction of an annotated corpus, the selection of features and their representations, and the choice of machine learning algorithms. In the following, we present the research background and related work of each component.

2.1. Constructing annotated corpora from MEDLINE

One step towards constructing annotated corpora from MEDLINE is to select a subset of MEDLINE abstracts. There are different ways to obtain such subset. One approach is to use keyword search. For example, abstracts selected for the GENIA corpus were retrieved from MEDLINE using three MeSH terms, "human", "blood cell" and "transcription factor" [4]. An alternative way to obtain a subset is to exploit the use of existing biomedical databases. For example, in order to construct an annotated corpus for the Interaction Article Subtask at the second BioCreative workshop, contents of two existing interaction databases, namely IntAct and MINT, have been exploited [5]. After deriving such subset, domain experts can manually annotate them.

2.2. Feature representation/selection

In order to use machine learning methods, each document needs to be transformed into a feature representation, which is usually a feature vector. Commonly, features are based on words appearing in the document. Various feature selection techniques have been explored to overcome the high-dimensionality of word-based features [6, 7], e.g., Term Frequency (TF), TF * Inverse Document Frequency (IDF), Information Gain (IG), Mutual Information (MI), or chi-square statistics. In this paper, we explored IG for feature selection. IG represents the quantity of information in a feature with regard to class prediction on the base of presence/absence of the feature in a document. Let $\{c_i\}_{i=1}^{m}$ be a set of categories to be predicted. Then IG of feature *w* in a document collection is defined as follows:

$$G(w) = E - E_1 - E_2,$$

$$E = -\sum_{i=1}^{m} P(c_i)\log_2 P(c_i),$$

$$E_1 = -P(w)\sum_{i=1}^{m} P(c_i \mid w)\log_2 P(c_i \mid w),$$

$$E_2 = -P(\overline{w})\sum_{i=1}^{m} P(c_i \mid \overline{w})\log_2 P(c_i \mid \overline{w}),$$

where *E* is the entropy of the document collection; *m* represents the number of categories; $P(c) = \frac{N_c}{N}$ is occurrence probability of category c, where N represents the number of documents and N_c is the file numbers of class c; $P(w) = \frac{N_w}{N}$ and $P(\overline{w}) = \frac{N_{\overline{w}}}{N}$ are occurrence probabilities of presence and absence of *w*, N_w and $N_{\overline{w}}$ are the file numbers of including and not including feature w in the document collection; and finally $P(c \mid w) = \frac{N_{wc}}{N_w}$ and $P(c \mid \overline{w}) = \frac{N_{\overline{wc}}}{N_{\overline{w}}}$ are occurrence conditional probability of the category c on occurrence and absence of term *w*, where N_{wc} and $N_{\overline{wc}}$ are the file numbers of including and not including term *w* in class *c*[8]. It is assumed that the larger the IG value of a term is, the more important the term is in classifying documents.

2.3. Machine learning algorithms

A growing number of statistical and probabilistic machine learning algorithms have been applied to document classification, including K nearest neighbor, Bayesian approaches, decision trees, symbolic rule learning, and neural networks [9-12]. Here, we chose Support Vector Machines (SVMs), a supervised learning algorithm proposed by Vladimir Vapnik and his co-workers [13, 14]. It has been widely used for text mining and achieved promising results. Given a training set with n class-labeled instances, (x_1, y_1), (x_2, y_2), ..., (x_i, y_i), …, (x_n, y_n), where x_i is a feature vector for the i-th instance and $y_i \in \{+1, -1\}$ indicates the class, an SVM classifier learns a hyper-plane as a decision boundary in the feature space. The class of an unlabelled instance x is determined by on which side of the hyperplane x lies. The purpose of training SVM classifiers is to find a hyperplane that has the maximum margin to separate the two classes [16-18].

3. Method

Figure 1 illustrates the overall data flow of the classification system. It consists of several steps including i) generating annotated MEDLINE abstracts, where each abstract was annotated either positive or negative (e.g., +1 or -1) based on its relevance to host-pathogen protein-protein interactions (PH-PPI), ii) conducting machine learning experiments to evaluate different kinds of feature representations and feature selection methods, and iii) implementing a system that assigns confidence scores to abstracts based on their PH-PPI relevance.

3.1. Generation of an annotated corpus

The annotated corpus was generated from two different sources. One was from UniProtKB database where the PH-PPI information is annotated for the protein entries and the relevant MEDLINE abstracts are cited. If a cited abstract contains an interaction pair consisting of one host protein and one pathogen protein, it is considered as positive. The other source was from PubMed, from which a set of MEDLINE abstracts was retrieved using keyword searches. Two domain experts reviewed and manually annotated this set, and categorized the abstracts as positive or negative. Additionally, for positive abstracts sentences describing the interactions were highlighted.

3.2. Machine learning

Instead of classifying a document as PH-PPI relevant or not, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. We defined two machine learning tasks. One task is at abstract level (ALT), which uses the abstracts to build a system to rank a set of abstracts according to their PH-PPI relevance. The other is on sentence level (SLT) which ranks all sentences in abstracts by considering titles and highlighted sentences in positive abstracts as positive and all sentences in negative abstracts as negative. The ranking of a set of abstracts can then be obtained according to the rank of the most relevant sentence in an abstract.

3.2.1. Feature representation/selection

We normalized the text by changing nouns in plural forms into singular forms, verbs in past tense into present tense, and replacing nouns and adjectives by their corresponding verbs based on the SPECIALIST lexicon, a component in the Unified Medical Language System (UMLS). We also replaced punctuation marks with spaces and changed

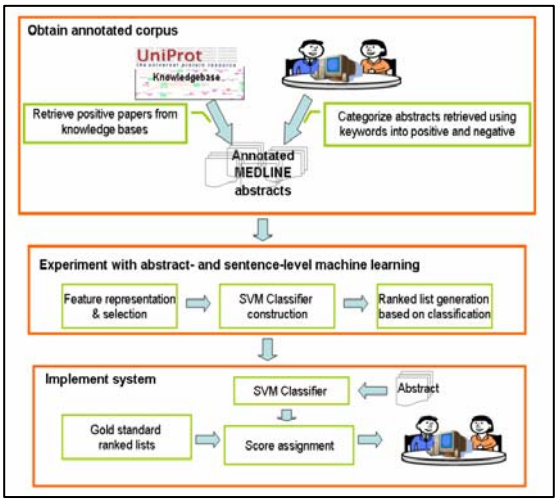


Figure 1. Overall architecture of the study.

uppercase letters to lowercase letters.

After normalization, we used unigrams and bigrams as features, and the frequencies of unigrams and bigrams as their corresponding feature values. To reduce the dimensionality of the feature space, we used information gain to select features with high IG values. Note that we did not remove features that are stop or rare words in this work.

3.2.2. Machine learning algorithms

We used the SVM light package and chose a linear function as the kernel [13]. We also experimented with other types of kernels such as polynomial or radial basis function (RBF), but observed no performance improvement.

3.2.3. Experiments

The experiments were designed to i) compare IG feature selection (IG-FS) with no feature selection (NO-FS), and ii) compare ALT and SLT. We used 100 runs of 10-fold cross validation. For each run, the same 10 fold partitions were used for the following four settings: (IG-FS, ALT), (IG-FS, SLT), (NO-FS, ALT), and (NO-FS, SLT). For each setting, we obtained a ranked list consisting of abstracts in the annotated corpus ranked according to the results of the 10-fold cross validation experiment. The performance was then measured using true positive rate (TPR): given rank threshold P and ranked list L, TPR(P, L) is defined as the ratio of the number of true positives ranked as top P in L to P. We selected 18 different rank thresholds: from 10 to 90 (incremented by 10) and from 100 to 500 (incremented by 50). In case of IG-FS, we set 20 IG thresholds: 0 to 0.0009

(incremented by 0.0001) and from 0.001 to 0.01 (incremented by 0.001). For each IG threshold, we ignored all features with IG values less than the threshold when constructing the systems. The average TPR of 100 runs for each setting was computed to compare the performance. Confidence intervals at 95% Confidence Level were also computed [15].

3.3. System implementation

As we have discussed, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. In order to judge the PH-PPI relevance for any given abstract, we used the following method:

i) obtain N score lists by executing N runs of 10-fold cross validation using the corpus as described in Section 3.2.3 where scores were ones assigned by SVM classifiers,

ii) build a SVM classifier C with all instances in the corpus,

iii) for a new abstract, use classifier C to obtain score S,

iv) for each score list that was obtained in i) compute the percentage of instances that are positive among the instances with scores larger than S, and

v) average the above percentage over N score lists and display the percentage as the relevance score. The higher the score, the more relevant the abstract.

To test the effectiveness of the proposed method, we used one run of 10-fold cross validation and measured TPRs for a given relevance score threshold.

4. Result and discussion

Most pathogen protein-protein interaction (PPI) information annotated in knowledgebases is for viral proteins or PPI within bacteria. We obtained less than 50 positive abstracts on specific bacterial pathogen-human host PPI from knowledge bases such as UniProtKB/Swiss-Prot and IntAct, Bru cella Bioinformatics Portal (BBP). Using keywords "bacterial", "host", "pathogen", and "interaction", we retrieved around 214,000 abstracts, and we obtained 1,225 negative abstracts and 99 positive abstracts after manual annotation. Merging the two sets, the annotated corpus consists of 1,225 negative abstracts and 135 positive ones.

Figure 2 shows the relationship between IG threshold and TPR averaged over 100 runs. The IG threshold of 0 corresponds to no feature selection (NO-FS). From Figure 2, we can see that for IG

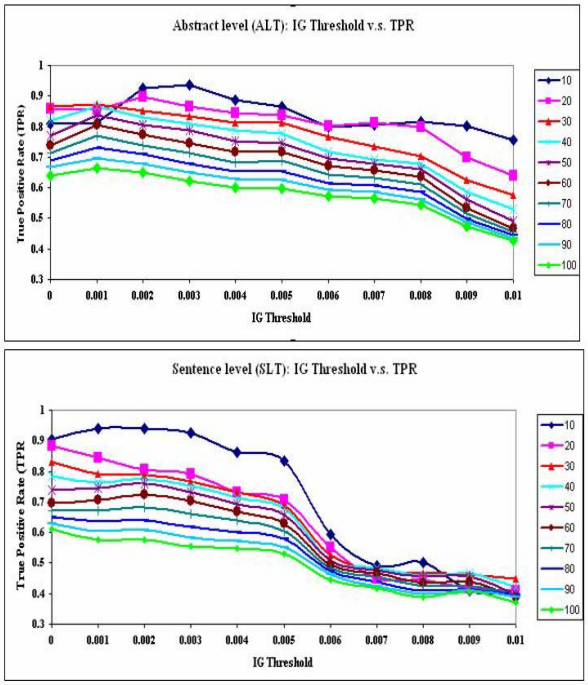


Figure 2. The relationship between IG threshold and TPR averaged over 100 runs in (IG-TF, ALT) and (IG-FS, SLT).

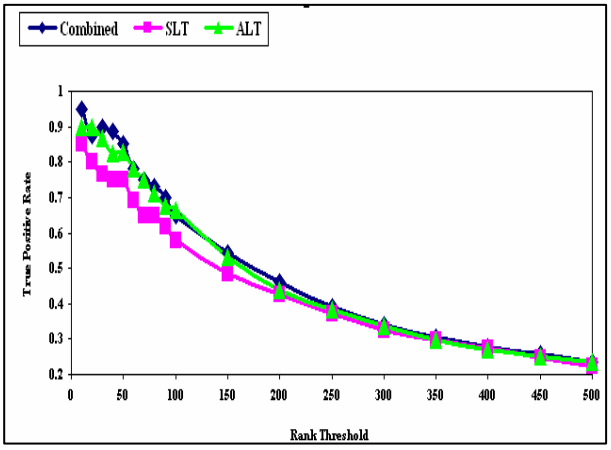


Figure 3. Combination result of (IG-FS, ALT)-0.001 and (IG-FS, SLT)-0.001.

thresholds between 0.001 and 0.005, the TPRs are comparable to the one without feature selection (i.e. NO-FS). However, the number of features used for classifiers with feature selection decreases dramatically. For example, in (IG-FS, ALT) with threshold 0.002 and (IG-FS, SLT) with threshold 0.001, the number of features after feature selection is reduced to only 10% (around 10,000) of the original (over 100,000).

Table 1. The detailed TPRs with the corresponding 95% confidence intervals computed from 100 runs for (IG-FS, ALT), (IG-FS, SLT), (NO-FS, ALT) with IG threshold 0.002, and (NO-FS, SLT) with IG threshold 0.001. RT stands for rank threshold.

RT	NO-FS		IG-FS	
	ALT	SLT	ALT(0.002)	SLT (0.001)
10	0.81 (0.794, 0.827)	0.905 (0.899, 0.911)	0.926 (0.915,0.937)	0.941 (0.930,0.952)
20	0.857 (0.852, 0.862)	0.883 (0.875, 0.891)	0.898 (0.890,0.906)	0.844 (0.834,0.854)
30	0.867 (0.862, 0.873)	0.832 (0.825, 0.839)	0.852 (0.845,0.859)	0.79 (0.782, 0.798)
40	0.819 (0.812, 0.826)	0.786 (0.779, 0.793)	0.83 (0.823,0.837)	0.764 (0.757, 0.771)
50	0.768 (0.762, 0.774)	0.737 (0.731, 0.744)	0.807 (0.801,0.813)	0.745 (0.739, 0.751)
60	0.74 (0.735, 0.745)	0.697 (0.692, 0.702)	0.775 (0.770, 0.780)	0.706 (0.700, 0.712)
70	0.715 (0.710, 0.720)	0.67 (0.665, 0.675)	0.738 (0.733, 0.743)	0.67 (0.665, 0.675)
80	0.69 (0.686, 0.694)	0.65 (0.646, 0.654)	0.71 (0.705, 0.715)	0.637 (0.633, 0.642)
90	0.666 (0.662, 0.67)	0.629 (0.625, 0.633)	0.679 (0.674, 0.684)	0.604 (0.600, 0.608)
100	0.639 (0.635, 0.643)	0.611 (0.6068,0.6152)	0.649 (0.645, 0.653)	0.577 (0.574, 0.581)
150	0.515 (0.513, 0.517)	0.514 (0.511, 0.517)	0.522 (0.519, 0.525)	0.491 (0.488,0.494)
200	0.431 (0.429, 0.433)	0.431 (0.429, 0.433)	0.438 (0.436, 0.440)	0.429 (0.427, 0.431)
250	0.377 (0.376, 0.379)	0.371 (0.369, 0.373)	0.378 (0.376, 0.380)	0.379 (0.377, 0.381)
300	0.336 (0.335, 0.337)	0.33 (0.329, 0.331)	0.334 (0.332, 0.336)	0.336 (0.334, 0.338)
350	0.303 (0.302, 0.304)	0.301 (0.300, 0.302)	0.3 (0.299, 0.301)	0.301 (0.300, 0.302)
400	0.278 (0.277, 0.279)	0.276 (0.275, 0.277)	0.273 (0.272, 0.274)	0.272 (0.271, 0.273)
450	0.257 (0.256, 0.258)	0.255 (0.254, 0.256)	0.251 (0.250, 0.252)	0.249 (0.248, 0.250)
500	0.238 (0.237, 0.239)	0.236 (0.235, 0.237)	0.232 (0.231, 0.233)	0.229 (0.228, 0.230)

Table 1 shows the detailed results of four settings: (NO-FS, ALT), (NO-FS, SLT), (IG-FS, ALT) with IG threshold 0.002, and (IG-FS, SLT) with IG threshold 0.001. For example, among top 50 abstracts, there are 76.8%, 73.7%, 80.7%, and 74.5% of the abstracts are positive for (NO-FS, ALT), (NO-FS, SLT), (IG-FS, ALT), and (IG-FS, LT), respectively. The average TPRs usually decrease when the rank thresholds increase. The performance of sentence-level systems is comparable to that of abstract-level systems when the rank threshold is small (e.g., 10 or 20). When the rank threshold (e.g., > 20) is large, abstract-level systems tend to perform better.

Table 2 shows the performance of the true positive rate when implementing the system using (IG-FS, ALT) with IG threshold 0.002 and the number of runs as 5. Given a relevance score threshold 0.5, the true positive rate is 50.7% which indicates that if an abstract receives a relevance score of larger than 0.5, the chance of the abstract to be positive is 50.7%.

Table 2. The performance of the implementation.

Threshold	Total	Positive	TPR
0	1,360	135	0.099
0.1	1,185	118	0.099
0.2	519	106	0.204
0.3	304	93	0.306
0.4	207	82	0.396
0.5	136	69	0.507
0.6	96	63	0.656
0.7	69	52	0.754
0.8	41	30	0.732
0.9	8	7	0.875

Even sentence-level systems perform inferior to abstract-level systems, but one advantage of them is that sentences describing protein interactions are automatically highlighted. We can highlight sentences

(and titles) yielding the highest rank among sentences within the abstract when presenting the results to end-users. For example, for (IG-FS, SLT) with IG threshold 0.001, the average number of positive abstracts is 17 (or 37) among the top 20 (or 50) abstracts. Among those positive abstracts, an average of 13 (or 26) abstracts have the highlighted sentences ranked as the highest among all sentences in the corresponding abstract by the sentence-level systems, and an average of 16 (or 33) abstracts have the highlighted sentences ranked as the highest or the second highest.

We also noticed that sentence-level systems and abstract-level systems behave differently. The finding is consistent with the work of Ding et al where different text units (e.g., abstracts, sentences, or phrases) were investigated for information retrieval [16]. Given rank threshold 10, and IG threshold 0.001, the average number of overlapped true positives between sentence-level and abstract-level systems is around 4. We checked the combination of sentence-level and abstract-level systems by averaging the ranks of sentence-level and abstract-level. Figure 3 shows the result. There is some improvement of the performance after combination.

5. Conclusion

We have reported a study of constructing an automated system that can detect the host pathogen protein-protein interaction relevance of MEDLINE abstracts. The results indicated that feature selection can reduce the number of features at least 10 folds with no or little sacrifice of performance. Additionally, the majority of the highlighted sentences are ranked as the first or second among all sentences in the corresponding abstracts. We conclude that automated systems can be built for retrieving abstracts and highlighting sentences based on their relevance to host pathogen protein-protein interaction.

6. Acknowledgements

This work was supported by US Army TATRC #W81XWH0720112 and NSF IIS-0639092.

References

[1] CG Zhang, BA Chromy and SL McCutchen-Maloney. Host–pathogen interactions: a proteomic view. *Expert Review of Proteomics*, 2(2):187-202, 2005.

[2] K Nomura, S DebRoy, YH Lee, N Pumplin, J Jones and SY He. A Bacterial Virulence Protein Suppresses Host Innate Immunity to Cause Plant Disease. *Science*, 313 (5784): 220-223, 2006.

[3] CH Wu, R Apweiler, A Bairoch, DA Natale, WC Barker, B Boeckmann, S Ferro, E Gasteiger, H Huang, R Lopez, M Magrane, M J Martin, R Mazumder, C O' Donovan, N Redaschi and B aris Suzek The Universal Protein Resource (UniProt): an expanding universe of protein information. *Nucleic Acids Research* 2006.

[4] J.-D. Kim, T. Ohta, Y. Tateisi and J. Tsujii. GENIA corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*. 19(Suppl. 1): i180-i182, 2003.

[5] M Krallinger and A Valencia. Evaluating the Detection and Ranking of Protein Interaction Relevant Articles: the BioCreative Challenge Interaction Article Sub-task (IAS). In *Proceedings of the Second BioCreative Challenge Evaluation Workshop*, pages 29-39, 2007.

[6] Y Yang and J Pederson. A Comparative Study on Feature Selection in Text Categorization". *Proceedings of the fourteenth International Conference on Machine Learning*, Pages 412-420, 1997.

[7] E C antú-Paz, S News am and C Kam ath. Feature selection in scientific applications. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, Pages 788 – 793, 2004.

[8] K Machová and A Szaboová. Statistical Methods in Key Words Generation from Text Documents. *In 5th Slovakian-Hungarian Joint Symposium on Applied Machine Intelligence and Informatics*, pages 435-446, 2007.

[9] Y Yang and CG Chute. An example-based mapping method for text categorization and retrieval. *ACM Transactions on Information Systems (TOIS)*, vol 12, Pages 252 – 277, 1994.

[10] DD Lewis, M Ringuette. A comparison of two learning algorithms for text categorization. In *Proceedings of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval*, pages 81-93, 1994.

[11] WW Cohen and Y Singer. Context-Sensitive Learning Methods for Text Categorization. *ACM Transactions on Information Systems (TOIS)*, Vol 17, Pages 141 – 173, 1999.

[12] ED Wiener, JO Pedersen and AS Weigend. A neural network approach to topic spotting. In *Proceedings of SDAIR-95, 4th Annual Symposium on Document Analysis and Information Retrieval*, Pages 317-332, 1995.

[13] VN Vapnik. Statistical Learning theory[M].1998.

[14] HF Liu, C Wu. A Study of Text Categorization for Model Organism Databases. *HLT-NAACL 2004 Workshop: Biolink 2004, Linking Biological Literature, Ontologies and Databases*, pages 25-32, 2004.

[15] U Hahn, M Romacker and S Schulz. Creating knowledge repositories from biomedical reports: the MEDSYNDIKATE text mining system. *Pac Symp Biocomput*, pages 338-349, 2002.

[16] J Ding, D Berleant, D Nettleton, and E Wurtele. Mining MEDLINE: abstracts, sentences, or phrases? *Pac Symp Biocomput*, pages 326-337, 2002.

Document classification for mining host pathogen protein-protein interactions

Lanlan Yin[a,2], Guixian Xu[b,c,a,2], Manabu Torii[d], Zhendong Niu[b], Cathy Wu[e], Zhangzhi Hu[f], Hongfang Liu[a,1]

[a]*Department of Biostatistics, Bioinformatics, and Biomathematics, Georgetown University, Washington DC, USA*
[b]*School of Computer Science and Technology, Beijing Institute of Technology, Beijing, China*
[c]*School of Information Engineering, Minzu University of China, Beijing, China*
[d]*Imaging Science and Information Systems Center, Georgetown University Medical Center, Washington DC, USA*
[e]*Protein Information Resources, Georgetown University Medical Center, Washington DC, USA*
[f]*Department of Oncology, Georgetown University Medical Center, Washington DC, USA*

Summary

Objective: Scientific findings regarding human pathogens and their host responses are buried in the growing volume of biomedical literature and there is an urgent need to mine information pertaining to pathogenesis-related proteins especially host pathogen protein-protein interactions (HP-PPIs) from literature.

Methods: In this paper, we report our exploration of developing an automated system to identify MEDLINE abstracts referring to HP-PPIs. An annotated corpus consisting of 1360 MEDLINE abstracts was generated. With this corpus, we developed and evaluated document classification systems using support vector machines (SVMs). We also investigated the effects of three feature selection methods (information gain, mutual information, and χ^2 test). The performance was measured using Normalized Discounted Cumulative Gain (NDCG) and Positive Predictive Value (PPV) and all measures were obtained through 10-fold cross validation.

[1]Corresponding Author: Building D, Room 180, Georgetown University, 4000 Reservoir Rd, NW Washington, DC 20007, USA. Phone: (202) 687-7933. Fax: (202) 687-2581.
[2]Equal contribution to the work

Results: NDCG measures for classification systems using all features or a subset of features selected using information gain and χ^2 range from 0.83 to 0.89 while classification systems built based on features selected using mutual information had relatively lower NDCG measures. The classification system achieved a PPV of 50.7% for the top 10% ranked documents comparing to a baseline PPV of 10.0%.

Conclusions: Our results indicate that document classification systems can be constructed to efficiently retrieve HP-PPI related documents. Feature selection was effective in reducing the dimensionality of features to build a compact system.

Key words: document classification, host pathogen protein-protein interaction, feature selection, literature mining

---

1. Introduction

The causative agents of infectious diseases consist of a great diversity of agents including bacteria, viruses, fungi, helminthes and protozoa. Because of the development of new molecular biology assays, there has been continuing progress in the study of pathogenicity mechanism. Meanwhile, due to the heightened concern about bioterrorism and emerging/reemerging infectious diseases, there have been major initiatives for large-scale genomic and proteomic projects to study the basic biology and disease-causing mechanisms of human pathogens [1, 2]. As a result, a flood of molecular data is being generated, but important scientific discoveries regarding these pathogens and their host responses are often buried under the increasing volume of biomedical literature. It was reported that the growth of peer-reviewed literature in MEDLINE is exponential [3]. With this volume of publication, it is very difficult or even impossible for biologists to find or assimilate the relevant publications of pathogenicity. To effectively manage the knowledge of pathogens and to better understand the pathogens, an automated text mining system that can extract pathogen related information from the scientific literature is highly desired.

In this paper, we focus on the development of an automated text mining system to identify research articles describing host pathogen protein-protein interactions (HP-PPIs). We focus on pathogens that are bacteria. By reviewing thousands of documents in MEDLINE, we constructed a corpus consist-

ing of 1360 abstracts where 135 abstracts are HP-PPI relevant (i.e., positive) and the remaining are not HP-PPI relevant (i.e., negative). The corpus was then used to train a machine learning classifier to identify HP-PPI related articles where samples are abstracts and features are words or phrases in the abstracts. Three feature selection methods, information gain (IG), χ^2 test, and specific mutual information (SI) were compared for reducing the high dimensionality of the feature space.

2. Background and related work

The task considered in this study is a special case of identifying papers that describe protein-protein interactions (PPIs). There are several components in developing an automated literature mining system, including the construction of an annotated corpus, the selection of features and their representations, and the choice of machine learning algorithms. In the following, we present the research background and related work of each component.

2.1. Construction of annotated corpora from MEDLINE

One step towards constructing annotated corpora from MEDLINE is to select a subset of MEDLINE abstracts. There are different ways to obtain such subset. One approach is to use keyword search. For example, abstracts selected for the GENIA corpus were retrieved from MEDLINE using three MeSH terms, "human", "blood cell" and "transcription factor" [4]. An alternative way to obtain a subset is to exploit the use of existing biomedical databases. For example, in order to construct an annotated corpus for the Interaction Article Subtask at the second BioCreative workshop, contents of two existing interaction databases, namely IntAct and MINT, have been exploited [5]. After deriving such subset, domain experts can manually annotate them.

2.2. Feature representation/selection

In order to use machine learning methods, usually each document needs to be transformed into a feature vector. Commonly, features are based on words appearing in the document. Various feature selection techniques have been explored to overcome the high-dimensionality of word-based features. In this paper, three widely used feature selection methods, information gain (IG), specific mutual information (SI), and χ^2 test, were applied and compared.

IG represents the quantity of information in a feature with regard to class prediction on the basis of presence/absence of the feature in a document. Let $\{c_i\}_{i=1}^m$ be a set of categories to be predicted. Then the IG value of feature t in a document collection $IG(t)$ is defined as follows:

$$IG(t) = E - E_1 - E_2, \tag{1}$$

$$E = -\sum_{i=1}^{m} P(c_i) \log_2 P(c_i), \tag{2}$$

$$E_1 = -P(t) \sum_{i=1}^{m} P(c_i|t) \log_2 P(c_i|t), \tag{3}$$

$$E_2 = -P(\bar{t}) \sum_{i=1}^{m} P(c_i|\bar{t}) \log_2 P(c_i|\bar{t}), \tag{4}$$

where E is the entropy of the document collection; $P(c)$ is the occurrence probability of category c; $P(t)$ and $P(\bar{t})$ are the occurrence probabilities of presence and absence of t; and finally $P(c|t)$ and $P(c|\bar{t})$ are the conditional probabilities of the occurrence of category c with or without feature t. The larger $IG(t)$ is, the more important t is. By calculating IG value for each variable appearing in the abstracts, a rank list for all the variables can be obtained. Given a threshold value, features with IG values ranked high are selected to build classifiers.

In information theory, the SI of two random variables has been used to describe the mutual dependence of the two variables. In text mining, the SI of feature t in category c, $SI(t, c)$, can be defined as:

$$SI(t, c) = \log \frac{p(t, c)}{p(t)p(c)}, \tag{5}$$

where $p(t, c)$ is the joint occurrence probability of t and c; and $p(t)$ and $p(c)$ are occurrence probabilities of t and c, respectively. Then the mutual information of t, $MI(t)$, can be defined as [6]:

$$MI(t) = \sum_{i=1}^{m} p(t, c_i) SI(t, c_i) + \sum_{i=1}^{m} p(\bar{t}, c_i) SI(\bar{t}, c_i) \tag{6}$$

The definition here yields the equivalence of $MI(t)$ and $IG(t)$ [7]. To distinguish from IG, Yang and Pedersen computed SI only based on the presence of

a specific term. The feature value of feature t was defined in two alternative ways [7]:

$$MI\_MAX(t) = \max_{i=1}^{m}\{SI(t, c_i)\}, \tag{7}$$

$$MI\_AVG(t) = \sum_{i=1}^{m} p(c_i)SI(t, c_i). \tag{8}$$

Feature words were ranked accordingly and only the top-ranked features were used to build classifiers.

The third feature selection method we applied is χ^2 test which is commonly used to test the independence of two variables. Here, the two variables are feature t and document class c. The null hypothesis is that the occurrence of t and the occurrence of c are independent. The statistics of χ^2 is defined as:

$$\chi^2(c, t) = \sum_{ec\in\{0,1\}} \sum_{et\in\{0,1\}} \frac{(O_{et,ec} - E_{et,ec})^2}{E_{et,ec}}, \tag{9}$$

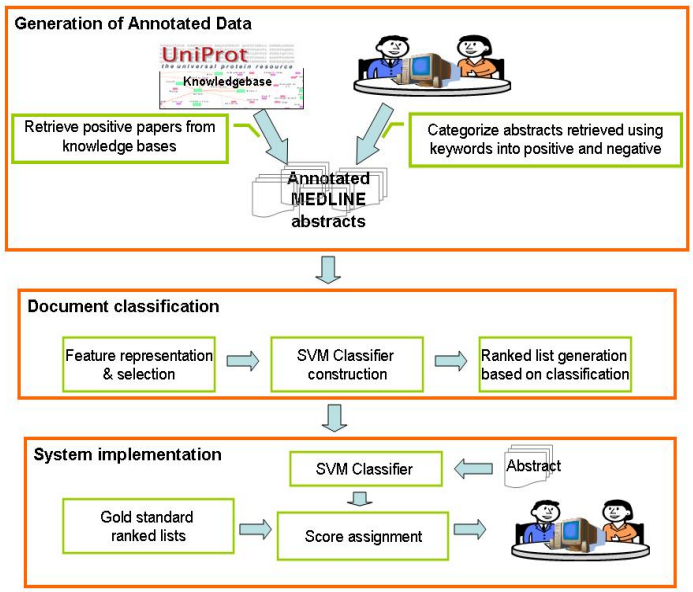
where χ^2 is the test statistic that asymptotically approaches a χ^2 distribution. O is an observed frequency; E is an expected (theoretical) frequency, asserted by the null hypothesis. Similarly, features are ranked with respect to their χ^2 scores, and the top-ranked features in are selected to train the classifier, since a high χ^2 score indicates that the hypothesis of independence between the feature and the class is incorrect.

3. Method and experimental design

Figure 1 illustrates the overall data flow of constructing the classification system. It consists of several steps:

- Generating an annotated MEDLINE corpus: each abstract was annotated either positive or negative based on its relevance to HP-PPI;
- Reducing the high dimensional feature space: three feature selection methods (IG, MI, and χ^2 test), and the resulting features were applied to train classifiers;
- Evaluating the performance: ten-fold cross-validation was used to evaluate the performance.

Figure 1: Experimental design



3.1. Annotated data generation

In order to gather HP-PPI related abstracts, two biomedical databases were investigated. First, a data file was downloaded from UniProtKB, where the HP-PPI information is annotated for the protein entries, and the relevant MEDLINE abstracts are cited. If a cited abstract contains HP-PPI information, it is considered as positive, while unrelated abstracts are labeled as negative. Second, a set of MEDLINE abstracts obtained by keyword searching were reviewed by two domain experts. The pathogen related and unrelated abstracts were tagged manually. Mining these two databases resulted in 135 positive abstracts and 1225 negative abstracts, with a total of 1360 samples.

3.2. Feature representation and selection

Each document was normalized by changing lexical variants to their base forms and replacing nouns and adjectives by their corresponding verbs based on the SPECIALIST lexicon, a component of the Unified Medical Language System (UMLS) [8]. We also replaced punctuation marks with spaces, and changed uppercase letters to lowercase letters. After normalization, we used uni-grams and bi-grams as features. An n-gram is a subsequence of n items

from a given sequence. Accordingly, our features included every single normalized word (uni-gram) in the corpus and every two neighboring normalized words (bi-gram) present in the corpus. The frequency of a uni- and bi-gram in each abstract was used as the feature value. Three feature selection methods introduced previously were applied. Additionally, for mutual information, we experimented with different document frequency thresholds where features with frequency lower than the given threshold were removed.

3.3. Document classification

A growing number of statistical and probabilistic machine learning algorithms have been applied to document classification, including K nearest neighbor, Bayesian approaches, decision trees, symbolic rule learning, and neural networks [9]. Here, we chose Support Vector Machines (SVMs), a supervised learning algorithm proposed by Vapnik and his co-workers [10, 11]. It has been widely used for text mining and achieved promising results. The purpose of training SVM classifiers is to find a hyperplane to separate the two classes with the maximum margin [10, 11]. SVMlight, by Joachims, is one of the most widely used SVM classification and regression packages. The algorithms used in SVMlight has scalable memory requirements and can handle problems with many thousands of support vectors efficiently [12, 13]. In the present project, we used the SVMlight package and chose the linear kernel. We also experimented with other types of kernels such as polynomial or radial basis function (RBF), but observed no performance improvement.

3.4. Performance evaluation

The performance was evaluated through 10-fold cross validation. In 10-fold cross validation, an annotated corpus is partitioned into 10 portions, and each portion is used to evaluate a classifier trained with the remaining 9 portions. Instead of traditional binary classification, for each run, we generated a rank list based on the classification scores.

The following metrics were used to measure the performance:

- Simplified Normalized Discounted Cumulative Gain (NDCG).

$$NDCG = Z_k \sum_{m=1}^{k} \frac{2^{R_m} - 1}{\log(1+m)} \tag{10}$$

  where Z_k is a normalization factor calculated to make it so that the NDCG of a perfect ranking at k is 1. R_m is the relevance of an abstract

to HP-PPI, either 1 (relevant) or 0 (irrelevant), m is the rank of the abstract in the final list, and k is the total number of the abstract [6]. The advantage of NDCG is that among the classifiers with same accuracy, the classifier which can rank the true positive literature higher will be awarded more.

- Receiver Operating Characteristic curve (ROC curve). This is a graphical plot of the true positive rate against the false positive rate for the different possible cut-points of a binary classifier system [14].

- Another measure used is the Positive Predictive Value (PPV) [15] which is the same as precision (i.e., the probability of predicted positives to be true positives) given a cut-point of a binary classifier system.

3.5. System implementation

As we have discussed, the machine learning task considered here is to rank a set of documents according to their PH-PPI relevance. In order to judge the PH-PPI relevance for any given abstract, we used the following method:

- obtain N score lists by executing N runs of 10-fold cross validation using the corpus where scores were ones assigned by SVM classifiers,

- build an SVM classifier C with all documents in the corpus,

- for a new abstract d, use classifier C to obtain a score $S(d)$ for d,

- for each score list that was obtained, compute the percentage of documents that are positive among the documents with scores larger than $S(d)$, and

- average the above percentage over N score lists and display the percentage as the relevance score. The higher the score, the more relevant the abstract.

To test the effectiveness of the proposed method, we used one run of 10-fold cross validation and measured PPVs for a given relevance score threshold.

4. Results and discussion

4.1. Document frequency for specific mutual information

Figures 2 and 3 display the performance of SVM classifiers on our corpus after using MI_MAX and MI_AVG as the feature selection method with different document frequency thresholds. In general, MI_MAX has better performance than MI_AVG. The classification results showed that the NDCG (normalized discounted cumulative gain) value of both MI_MAX and MI_AVG generally decreases as the number of features decreased, which can be explained by the smaller amount of information (fewer features) recruited by the classifier. However, the performance of MI_MAX was improved as the document frequency threshold increased. By setting the threshold of document frequency, low frequency terms with document frequency less than the threshold can be removed from the feature space. In our case, the NDCG of the classifier based on MI_MAX remains above 0.83 even with only 1000 feature terms if the document frequency threshold was no less than 3, while the NDCG of other classifiers with threshold of 1 and 2 was less than 0.82 with 3000 feature terms. Therefore, setting document frequency threshold is a crucial step for applying MI_MAX. But for MI_AVG, the performance was not improved by increasing the threshold of document frequency. To calculate the average mutual information for each term, a weight was assigned to each term for each class. Here, we use the occurrence probability of each class as the weight. Due to the imbalanced distribution of classes (only 10% documents are positive), the weight for the terms in positive abstracts would be 0.1, much lower than the terms in negative abstracts. Consequently, the informative terms in positive documents were swamped by the terms in negatives. In our project, the features of positive documents are more helpful in recognizing the pattern. Together, the poor performance of MI_AVG and the distinct characteristics from MI_MAX were caused by the bias towards low-frequency words and the bias towards words in negative abstracts.

4.2. Comparison of feature selection methods

Table 1 and Figure 4 show the comparison results of three feature selection methods. When there were more than 4000 feature terms, MI_MAX, IG, and CHI had similar performance. But as the number of features used in the classifier decreases to less than 4000, the performance of MI_MAX declines much faster than IG and CHI. The classifier curve goes to the minimum 0.769 if the classifier used 100 terms selected based on MI_MAX, while using

Figure 2: Performance of maximum mutual information (MI_MAX)

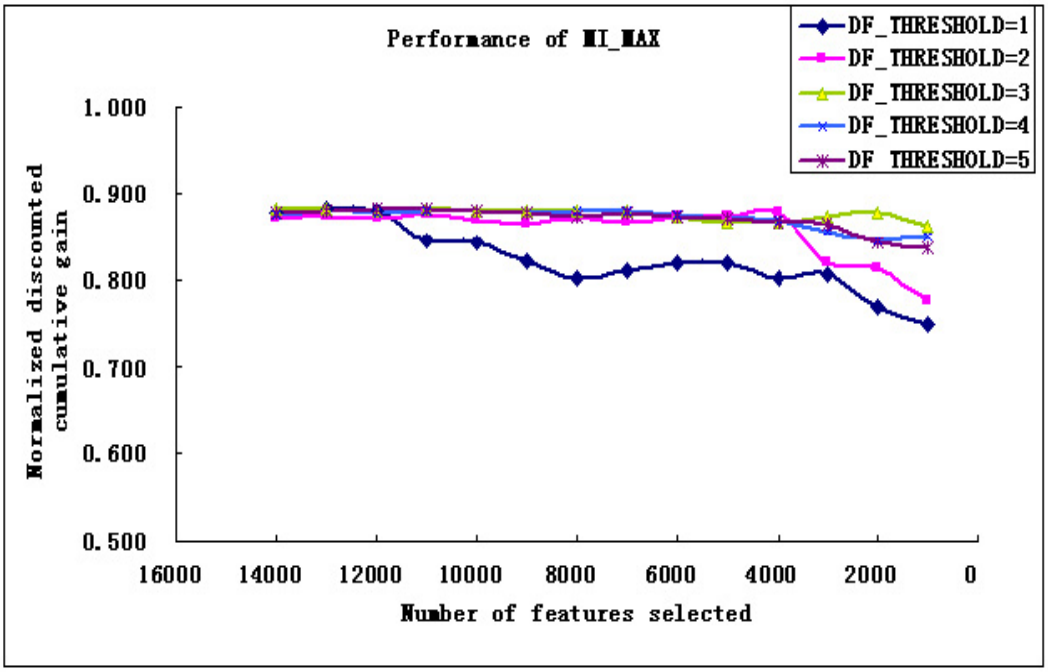
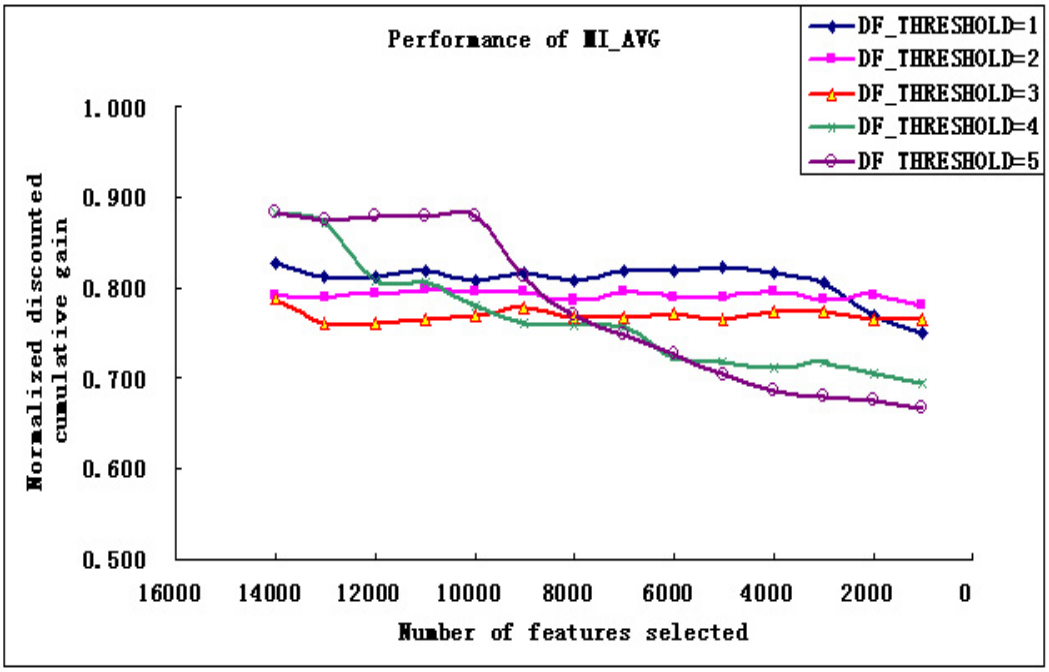


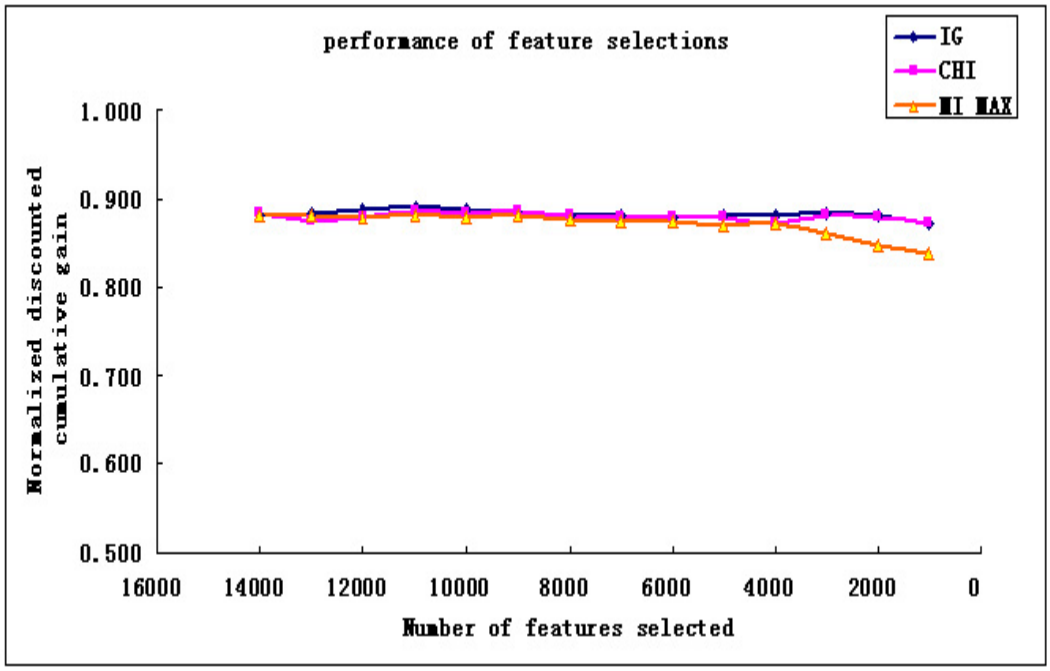
Figure 3: Performance of average mutual information (MI_AVG)



FEATURE	IG		CHI		MI_MAX	
NUMBER	AVG	SD	AVG	SD	AVG	SD
14,000	0.881	0.011	0.883	0.006	0.881	0.008
13,000	0.885	0.013	0.875	0.013	0.881	0.007
12,000	0.888	0.007	0.880	0.009	0.879	0.009
11,000	0.890	0.006	0.885	0.005	0.882	0.004
10,000	0.888	0.005	0.884	0.006	0.880	0.007
9,000	0.886	0.008	0.886	0.005	0.881	0.009
8,000	0.882	0.006	0.883	0.006	0.876	0.006
7,000	0.882	0.007	0.880	0.006	0.874	0.008
6,000	0.880	0.006	0.878	0.008	0.876	0.005
5,000	0.882	0.007	0.880	0.007	0.871	0.007
4,000	0.881	0.006	0.874	0.007	0.873	0.007
3,000	0.883	0.006	0.881	0.006	0.862	0.005
2,000	0.881	0.009	0.880	0.006	0.846	0.011
1,000	0.874	0.010	0.872	0.015	0.838	0.015
900	0.876	0.006	0.878	0.012	0.843	0.018
800	0.871	0.011	0.877	0.008	0.831	0.015
700	0.872	0.008	0.872	0.007	0.833	0.013
600	0.873	0.008	0.872	0.011	0.842	0.015
500	0.874	0.005	0.863	0.012	0.839	0.014
400	0.874	0.009	0.860	0.013	0.831	0.015
300	0.868	0.009	0.866	0.007	0.817	0.013
200	0.862	0.006	0.862	0.007	0.792	0.013
100	0.850	0.009	0.831	0.008	0.769	0.010

Table 1: Average NDCG of classifiers with feature selection

Figure 4: Comparison results of three feature selection methods



the same number of features selected from information gain or χ^2 test the classifier's NDCG is still above the line of 0.831, indicating MI_MAX does not have comparable performance to the other two methods: information gain and χ^2 test.

4.3. Comparison of systems with and without feature selection

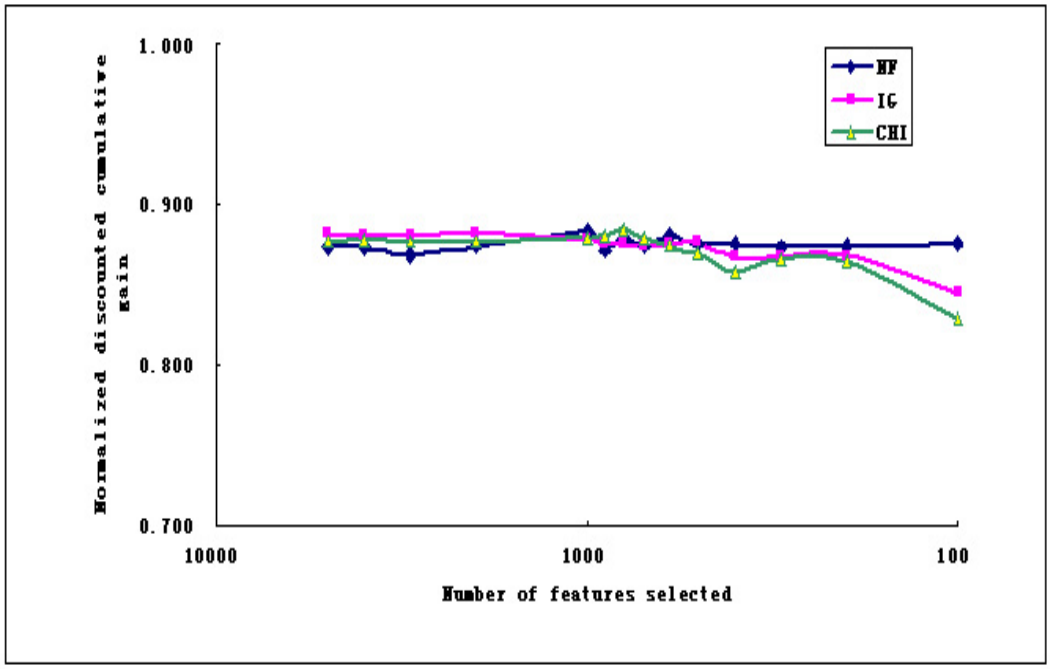
The overall performance of no feature selection, information gain, and χ^2 test is shown in Figure 5. We found that as the feature number became greater than 500, the performance of the IG and χ^2 test were comparable to that of no feature selection. Both IG and χ^2 test outperformed no feature selection by virtue of the much lower dimensional feature spaces they used. These two feature selection methods selected a small number of variables and then generated compact models.

In implementation, each abstract in the test dataset was assigned a score by an SVM classifier, and the abstracts were ordered by those scores. The higher the score, the more likely the abstract to be positive. Therefore, given a rank threshold N, the abstracts with rank above N were classified as positive abstracts, while the abstracts with lower rank were categorized as negative. Given a series of rank thresholds, ROC curves of different classifiers built upon no feature selection, information gain, and χ^2 test were shown in Figure 6. All three curves approach the left-hand border and then the top border of the ROC space, located far from the no-discrimination line, indicating competent classification capability. The χ^2 curve lies closer to the 45-degree diagonal of the ROC space, suggesting poor performance. Figure 7 shows the positive predictive value of three models at rank thresholds of 25, 50, 75, 100, and 135. A classifier with no feature selection gave the best performance among the three at each threshold because it utilized all uni-grams and bi-grams as features, thereby using as much information as possible from the samples. However, classifiers build upon information gain and χ^2 achieved comparable results with much lower cost. The number of term used was reduced to 2000 (a 98.3% reduction).

4.4. Evaluation of implementation

Table 2 shows the performance of PPVs when implementing the system using information gain (IG) threshold of 0.002 and the number of runs parameter is set to 5. Given a relevance score threshold 0.5, PPV is 50.7% which indicates that if an abstract receives relevance score higher than 0.5,

Figure 5: Performance of no feature selection vs. feature selection



Threshold	TOTAL	TP	PPV
0	1360	135	0.099
0.1	1185	118	0.099
0.2	519	106	0.204
0.3	304	93	0.306
0.4	207	82	0.396
0.5	136	69	0.507
0.6	96	63	0.656
0.7	69	52	0.754
0.8	41	30	0.732
0.9	8	7	0.875

Table 2: The performance of implementation

Figure 6: Receiver operating characteristic curve of different classifiers

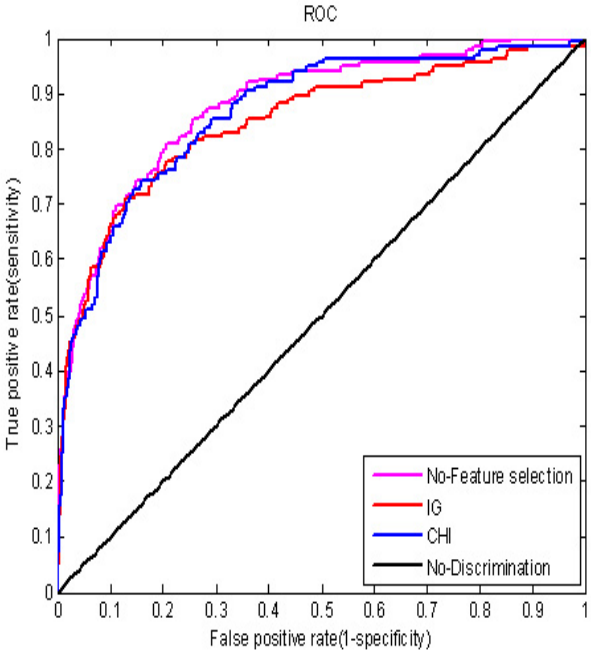
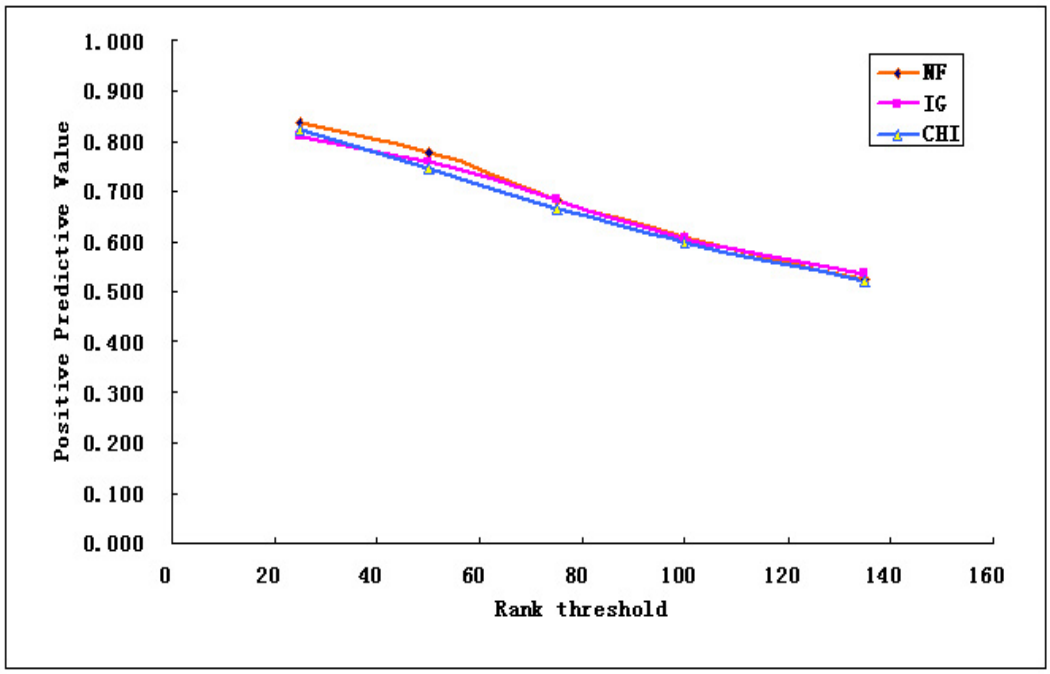


Figure 7: The positive predictive value of different classifiers



the probability of the abstract to be positive is 50.7%, which is much higher than the random chance to select positive abstracts (9.9%; 135 out of 1360).

5. Conclusions

In summary, we built a text mining system that retrieves MEDLINE abstracts pertaining to host-pathogen protein-protein interaction. We manually constructed a literature corpus consisting of 1360 Medline abstracts, where 135 are HP-PPI related and the remaining ones are HP-PPI unrelated. This corpus was used to build automated text categorization system that classifies MEDLINE abstracts as HP-PPI related or not. As a classification algorithm, SVM was used. In addition, three feature selection methods (IG, MI, and χ^2 test) were considered to reduce the high dimensionality of the feature space. Among them, IG and χ^2 test were found effective in reducing the dimensionality and, thus, in building a compact system. Our results indicate that an automated document classification system can help curators search and retrieve HP-PPI related biomedical literature.

References

[1] C.G. Zhang, B.A. Chromy, and S.L. McCutchen-Maloney. Host-pathogen interactions: a proteomic view. *Expert Rev. Proteomics*, 2(2):187–202, 2005.

[2] K. Nomura, S. DebRoy, Y.H. Lee, N. Pumplin, J. Jones, and S.Y. He. A bacterial virulence protein suppresses host innate immunity to cause plant disease. *Science*, 313(5784):220–223, 2006.

[3] L. Hunter and K.B. Cohen. Biomedical Language Processing: What's Beyond PubMed? *Molecular Cell*, 21(5):589–594, 2006.

[4] J.D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. GENIA corpus: a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(1):180–182, 2003.

[5] M. Krallinger, F. Leitner, C. Rodriguez-Penagos, and A. Valencia. Overview of the protein-protein interaction annotation extraction task of BioCreative II. *Genome Biology*, 9(Suppl 2):S4, 2008.

[6] C.D. Manning, P. Raghavan, and H. Schütze. *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA, 2008.

[7] Y. Yang and J.O. Pedersen. A comparative study on feature selection in text categorization. In D.H. Fisher, editor, *The Fourteenth International Conference on Machine Learning*, pages 412–420. Morgan Kaufmann Publishers, Inc., San Francisco, CA, USA, 1997.

[8] O. Bodenreider. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic Acids Research*, 32(Database Issue):D267, 2004.

[9] F. Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[10] T. Joachims. Text categorization with support vector machines: learning with many relevant features. In C Nedellec and C. Robveirol, editors, *Proceedings of ECML-98, 10th European Conference on Machine Learning*, pages 137–142. Springer Verlag, London, UK, 1998.

[11] V. Vapnik. *Statistical Learning Theory*. Wiley-Interscience, New York, NY, USA, 1998.

[12] T. Joachims. *Learning to classify text using support vector machines: methods, theory and algorithms*. Kluwer Academic Publishers Norwell, MA, USA, 2002.

[13] T. Joachims. *Making large-scale support vector machine learning practical, advances in kernel methods: support vector learning*. MIT Press, Cambridge, MA, USA, 1999.

[14] T. Fawcett and P.A. Flach. A response to webb and ting's on the application of roc analysis to predict classification performance under varying class distributions. *Machine Learning*, 58(1):33–38, 2005.

[15] D. Altman and J. Bland. Diagnostic tests 2: predictive values. *BMJ. British medical journal(International ed.)*, 309(6947), 1994.

iProLINK: A Framework for Linking Text Mining with Ontology and Systems Biology

Zhang-Zhi Hu[1*], K. Bretonnel Cohen[2], Lynette Hirschman[2], Alfonso Valencia[3], Hongfang Liu[4], Michelle G. Giglio[5], Cathy H. Wu[1]

[1]Protein Information Resource, [4]Department of Biostatistics, Biomathematics and Bioinformatics, Georgetown University Medical Center, Washington DC; [2]Information Technology Center, The MITRE Corporation; [3]Spanish National Cancer Research Centre (CNIO), Spain; [5]University of Maryland School of Medicine, Baltimore, MD

{zh9, hl224, wuc}@georgetown.edu; {kbcohen, lynette}@mitre.org; valencia@cnio.es; mgiglio@som.umaryland.edu

Abstract

The ever-increasing scientific literature and the exponential growth of large-scale molecular data have prompted active research in biological text mining to facilitate literature-based curation of molecular databases. Meanwhile, systems biology and bio-ontologies are emerging as critical tools in biological research where complex data in disparate resources are generated, integrated and analyzed. Both rely on literature for data annotation and analysis. The challenges facing us are to develop broadly utilized text mining tools and systems, and to bring together developer and user communities for system development and evaluation. We describe a framework for linking text mining tools with ontology and systems biology, extending from a previously developed text mining resource, iProLINK. We focus on molecular and ontological resources, including genes/proteins, protein-protein interaction (PPI), and Protein Ontology. The framework consists of two major components: a user interface for text mining of PPI from an integrated tool server and software modules to allow text mining outputs to be created, ranked, and used by the community. Use cases are presented for assessing the gaps and making recommendations for future development.

1. Introduction: current status of text mining as an enabling tool for biology

The biological literature represents the repository of biological knowledge. As biology becomes more dependent on information technology, there has been an explosion of computable resources and databases [1], e.g. GenBank, UniProt, model organism databases, and systems biology databases, e.g., Reactome, KEGG, that capture much of the structured information on sequence and functional data. It becomes critical to link these data sources to their associated context, e.g., experimental methods and evidence. Such information is largely buried in the literature and it has become prohibitively expensive for curators to keep up with its growth.

1.1. Text mining resource development

Most of the work in biomedical text mining over the past decade has focused on solving specific problems, often using task-tailored and private datasets, which were rarely reused. As more research groups began to make resources publicly available, there have been a number of projects, initiatives and organizations dedicated to building and providing access to biomedical text mining resources, such as those listed at the National Center for Text Mining at UK (http://www.nactem.ac.uk) and Text Mining Group at the Center for Computational Pharmacology (http://compbio.uchsc.edu/ccp/corpora).

Researchers at PIR have contributed to this effort by developing a literature mining resource, iProLINK, to support text mining and NLP research for bibliography mapping (references cited in a protein entry), annotation extraction, entity recognition and protein ontology development [2]. The data sources for bibliography mapping and feature evidence attribution include mapped citations and annotation-tagged literature corpora [3]. The data sources for entity recognition and ontology development include protein name dictionaries and protein name-tagged literature corpora along with tagging guidelines [4]. These curated corpora have been used for training and benchmarking text mining tools such as RLIMS-P, an information extraction tool for protein phosphorylation [5]. iProLINK also provides the online BioThesaurus, a large collection of gene/protein names with UniProt entry associations [6].

1.2. Text mining critical evaluations

* Corresponding author.

As the BioCreative [7, 8] and TREC Genomics track [9] evaluations have shown, common evaluations are important to create an active research community and to accelerate the research progress. There have been two BioCreative workshops to date, with 27 groups participating in the first [7], and 44 groups participating in the second [8]. These workshops have focused on tasks relevant to the biological curation community, including identification of gene mention (GM) and gene normalization (GN), and on more advanced tasks. For BioCreative I, the focus was on functional annotation, including linkage of evidence passages to support GO annotations for proteins in full text articles. For BioCreative II, the advanced task focused on extraction of protein-protein interaction (PPI) information, using "gold standard" data provided by the MINT and IntAct databases. The BioCreative evaluations have driven progress in biomedical text mining and have led to release of annotated data collections for further evaluation (http://BioCreative.sourceforge.net).

1.3. Text mining tool integration

It has been observed that "accurate and diverse" tools targeting the same application area can make a combination system outperform a single constituent tool [10, 11]. For example, Si et al. [12] combined systems that participated in the JNLPBA shared task (recognition of five types of entities in abstracts), and reported excellent performance using Conditional Random Fields (CRFs). Similarly [13, 14] reported results obtained by combining 21 systems from the BioCreative II GM task, and reported an F-measure over 90% using CRFs.

A major accomplishment of BioCreative II was the establishment of BioCreative MetaServer (BCMS, http://bcms.bioinfo.cnio.es/) [15], a prototype platform that combines text mining services from multiple groups, currently covers some major tasks from BioCreative II, including GM/GN, taxon classification and PPI identification, and provides annotations from 13 servers for the BioCreative corpus of MEDLINE abstracts.

1.4. Text mining standards development

Common standards for data exchange and tool integration are critical for text mining. Currently there is a lack of formal standards and candidates for de facto standards are not widely accepted at this time. The first concrete proposal for a data exchange standard for biomedical text processing was GPML, the GENIA Project Mark-up Language [16]. A corpus annotated in this format has been released in multiple revisions and has experienced significant acceptance in the text mining community [17], but tool producers have not embraced it as an output format. For the tool integration, there has been considerable amount of interest in the Unstructured Information Management Architecture (UIMA) [18-21], but it is not considered the de facto standard for tool integration yet. A meeting held in conjunction with the recent BioNLP 2008 workshop concluded that there was little hope for convergence on a common format in the near future, and that the best that could be hoped for at this time with respect to corpora and data exchange is that corpus builders produce formats that can be interconverted—no small feat in itself [22].

1.5. Motivation for a community framework

Even with advancements in tool and system development and the growing collaborative efforts of the text mining community, literature mining tools are still not broadly used by biological communities. Such a gap is partly due to intrinsic complexity of biological text for mining, and partly to the lack of close interactions between the text mining and the user communities, represented by biology researchers and curators.

BioCreative I and II focused on critical assessment of text mining tool performance on individual tasks involved in the overall molecular data curation process. The next step is to link these tools together to provide an environment that supports end users. The communities represented by biologists/curators and tool developers can be brought together by a common interface and through community workshops. In this paper, we describe an extended iProLINK framework that aims to link the three communities, allowing text mining tools to be evaluated and adopted by the broad communities. This work builds on four threads of research: the previous iProLINK text mining resource; BioCreative evaluations; tools and data resources developed under BioCreative, in particular the vision of a MetaServer to provide text mining services to users; and work at PIR focused on building a framework for the capture of PPI, including post-translational modifications (PTM). We present several case studies that illustrate the mutual benefit each community can gain from the others.

2. Linking text mining with ontology and systems biology: a basic framework

2.1. iProLINK framework

An overview of the iProLINK framework is shown in Figure 1. It contains two major components: text mining tools, and the interface that links the text mining to ontology and systems biology communities. Text mining tools are integrated into a metaserver that will generate text mining results, and the user interface will display ranked outputs (circle #1) and the visualized protein networks (#2) based on the output. The interface also allows users/curators to curate the text mining results (#1) and make assertions on the extracted knowledge. The curated information is used for or captured in ontologies (e.g. Protein Ontology) (#3) and

knowledgebases (#4), and is also saved in a curated literature corpus (#6) used for improving the text mining output ranking (#7) and for enhancing text mining tool development (#8). The systems biology data can also be used to help assertion of the text mining results (#5).

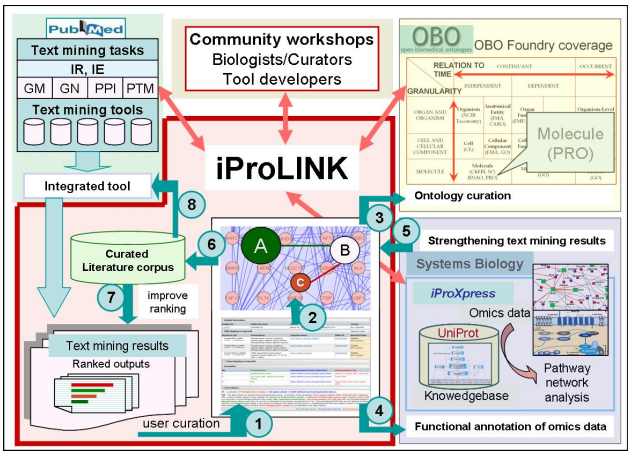


Figure 1. Overview of the iProLINK framework

2.2. Linking text mining, ontology and systems biology for protein-protein interactions

PPI generally refers to physical associations of two protein objects, stable or transient, such as in protein complexes or in signaling cascades. There are many types of PPIs; in this context, we define PPI as protein pairs with either direct or indirect associations such as through intermediate steps.

Text mining. The text mining tasks for iProLINK include integration of tools, presently covering gene or protein mention, gene or protein normalization, and information retrieval and extraction of PPI, including PTMs such as phosphorylation (an interaction between a protein substrate and a protein kinase). There are a number of tools for these tasks, including those participating in the BioCreative I and II challenge evaluations, and others such as RLIMS-P.

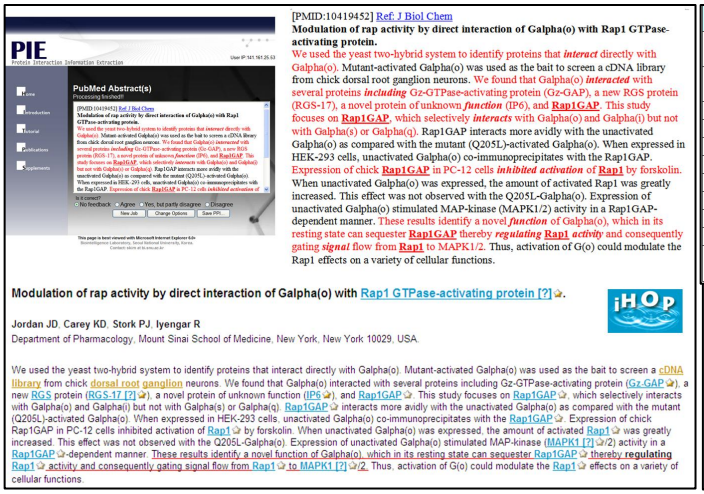
Ontology. Open Biological Ontologies (OBO) Foundry is a collaborative effort for coordinating various biological ontology development projects and for fostering common standards in OBO development [23]. The curation of the content of ontologies, especially those related to genes or proteins, e.g. specific splice or modified forms of gene products in Protein Ontology (PRO) [24], relies heavily on literature information. In particular, protein PTM and PPI text mining will help annotate protein nodes (terms) by identifying specific phosphorylated forms and adding PPI information as attributes to PRO forms.

Systems biology. Molecular databases represent structured knowledge of genes/proteins, such as UniProt, and biological pathway and PPI databases. Annotation of those databases and utilization of the annotations for large-scale omics data analysis are an integral part of systems biology, e.g., iProXpress, an expression analysis system for systems biology [25]. Text mining results can be used to infer or add more evidence to pathway and network analysis results derived from systems biology data; conversely, large-scale data can be used to support the text mining results of PPI information.

3. iProLINK use case analysis

3.1. PPI text mining for generation of protein networks

There are several PPI text mining tools, such as PIE [26] and iHOP [27], both as part of the BCMS. We use these two tools to illustrate PPI text mining results and how they can be used for generation of protein networks. As shown in Figure 2, the tools typically highlight or underline sentences containing the PPI, with protein pairs and words for relations highlighted (bold or colors). There are 11 pairs of PPI instances in this abstract, including the title. Most (8/11) are detected by one or the other tool, and most (9/11) are direct PPIs.



Protein 1	Protein 2	Tools
Galpha(o)	Rap1 GTPase-activating protein	n/a
Galpha(o)	Gz-GTPase-activating protein (Gz-GAP)	PIE
Galpha(o)	RGS protein (RGS-17)	PIE
Galpha(o)	IP6	PIE
Galpha(o)	Rap1GAP	PIE
Rap1GAP	Galpha(o)	PIE
Rap1GAP	Galpha(i)	PIE
Galpha(o)	MAP-kinase (MAPK1/2)	n/a
Galpha(o)	Rap1	n/a
Rap1	MAP-kinase (MAPK1/2)	iHOP
Rap1GAP	Rap1	PIE iHOP

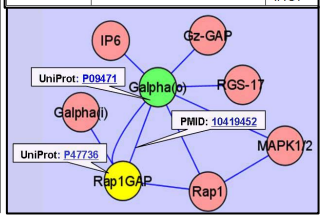


Figure 2. PPI text mining results for the construction of protein network

The visualized PPI network allows users to more efficiently mine proteins of interest and their interacting partners. Based on the binary relations (edge) between interacting partners (node), we used Cytoscape [28] to display these mined PPIs in a single protein network (Figure 2, lower right). It shows that Galpha(o) is a

major hub protein that interacts with six other proteins directly or indirectly. Rap1GAP is another important protein that interacts with three other proteins. The UniProt IDs for the protein nodes are displayed with mouse-over, and the text evidence for relations (edges) between protein nodes is also visualized by mouse-over (PMID in this case). The protein networks can also be built from multiple abstracts either through batch retrieval (section 3.3) or by gene/protein name searches. The latter would be a more useful feature in analyzing PPI of particular proteins based on PubMed searches.

The essential requirements for the interface in PPI text mining and protein network generation are to 1) provide ranking of PPI outputs based on scores or confidence levels for each protein pairs; 2) support user/curator feedback on the output ranking and content, and an ability to save the output in standard data formats compatible to other software tools such as Cytoscape and OBO editor; and 3) display the protein nodes and edges with weightings and evidence attributions.

3.2. PTM text mining for Protein Ontology form curation

The Protein Ontology is designed to describe the relationships of proteins and protein evolutionary classes, to delineate the multiple protein forms of a gene locus, and to interconnect existing ontologies [24]. Multiple protein forms include splice isoforms and various PTMs. Knowledge of protein splice forms and modifications are mostly embedded in the literature, thus text mining of such information greatly facilitates the curation of PRO nodes (terms) and relations. Protein phosphorylation is a common type of PTM, and proteins with distinct phosphorylated residue(s) represent unique protein forms. RLIMS-P is designed to extract the three protein phosphorylation objects: kinase, substrate and the phosphorylation sites/residues. The kinase and substrate interaction is a special case of PPI that can be mined by text mining tools, such as PIE. However, RLIMS-P also extracts phosphorylation sites, useful for PRO curation.

Figure 3 shows the output of the RLIMS-P extracted PPI and phosphorylation sites (PMID: 18003885), which can be directly used for curation of the protein node, RUNX1, a transcription factor. RLIMS-P outputs contain a summary table for the extracted PPI and evidence-tagged sentences in the abstract. One of the 11 isoforms, AML-1G, of human RUNX1 is described in PRO format as being phosphorylated at Ser 48, 303, and 424; the specific PTM type (phosphorylation at L-serine) is annotated using the PSI-MOD ontology (MOD:00046) (Figure 3). Experimental PPI information can also be used for annotating properties to protein forms in PRO, e.g., the associated functions of the phosphorylated form of RUNX1 in this paper can be annotated for AML-1G, e.g., "increases transactivation potency and stimulates cell proliferation". The RLIMS-P outputs need to be saved in standard formats, such as OWL or OBO, for protein network display and PRO curation.

Protein 1	Protein 2	Rel.	PMID	Tool
Galpha(o)	Rap1GAP	I	10419452	PIE
Cdc2	Rap1GAP	P	8048970	RLIMSP
Cdc2	BAD	P	12049737	RLIMSP PIE
BAD	14-3-3	I	12049737	RLIMSP PIE
			10949026	RLIMSP PIE
BAD	Bcl-xL	I	8929531	PIE

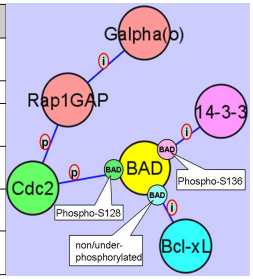


Figure 4. Text mining summary and network generation of PPI, including general "Interaction" (I) and protein phosphorlyation (P)

3.3. PPI text mining for systems biology

Systems biology data include gene/protein databases and large-scale omics data repositories. Annotation and analysis of systems biology data can benefit from PPI text mining. The protein network in Figure 2 contains the Rap1-MAPK pathway, which is modulated by Gα(o)-Rap1GAP interaction. Other papers describe the activation of Rap1GAP through phosphorylation by Cdc2 (CDK1), which also phosphorylates the BAD protein at distinct site (Ser128) (Figure 4). Interestingly, distinct forms of BAD interact with different partners.

When combining PPI mining results from Figure 2 and 4, a larger protein network can be generated, showing four highly-connected protein nodes—Gα(o), Rap1GAP, Cdc2 and BAD (Figure 5A). Compared to a pathway diagram

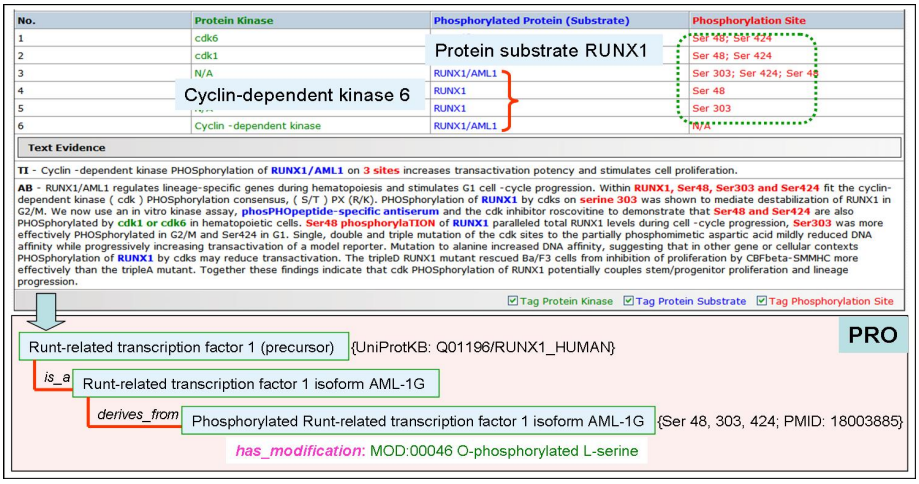


Figure 3. RLIMS-P text mining for Protein Ontology curation

based on the analysis of a proteomics dataset [29] (Figure 5B), this text mining-based PPI network graph not only provides literature evidence for the interactions shown in the pathway map (e.g., GNAO2-Rap1GAP, Rap1GAP-Rap1), but also reveals a missing interacting protein pair (Cdc2-Rap1GAP) in the pathway (red dashed arrow), as well as missing partners of BAD protein (14-3-3 and Bcl-xL).

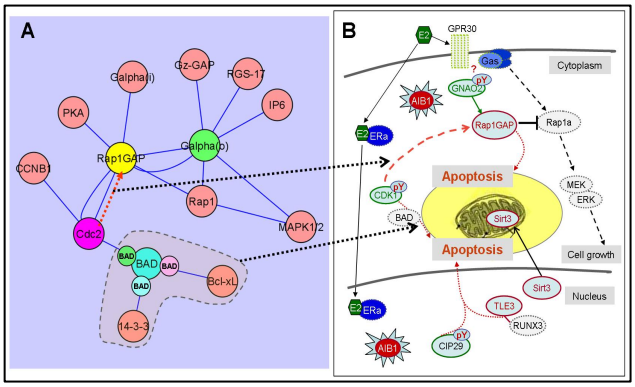


Figure 5. Text mining of PPI (A) for annotation and analysis of systems biology data (B)

3.4. PPI text mining supported by systems biology data

Systems biology data can also strengthen PPI text mining results. Figure 6 shows an example where PPI proteomic data from large-scale immunoprecipitation are linked to text mining results. The Sp1-p38 interaction from a proteomics experiment was deposited in IntAct, one of the PPI and pathway databases integrated into the iProXpress underlying data warehouse. This information supports the protein network derived from text mining, showing p38-Sp1 interaction and activation of filamin-A.

The display of protein networks will allow linkage of protein nodes to pathway maps or high-throughput PPI data from molecular databases. Alternatively, saved text mining outputs can also be integrated into users' pathway and network analysis pipeline.

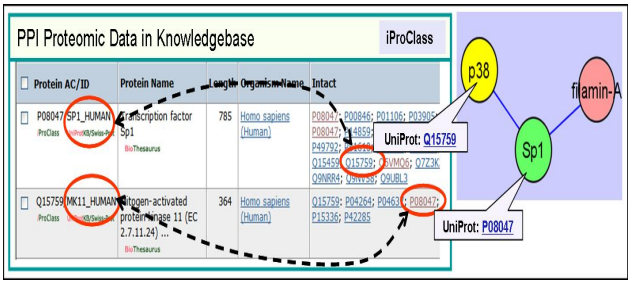


Figure 6. Systems biology data support the text mining of Sp1-p38 interaction (PMID: 12324467)

4. Future work

From above case studies, we have identified major gaps between the text mining and the ontology and systems biology communities that need to be addressed:

Standards development. Text mining standards include those of data exchange and tool integration. Tool integration involves issues such as process control and preserving state information as well as a mechanism for exchanging data. Standards must also support data exchange, including both syntactic standards (e.g., XML or SGML tags) and semantic standards – perhaps based on widely accepted biological resources, such as EntrezGene and UniProt.

User interface requirements. The web interface is a major component of the iProLINK framework for the communities. The new interface will allow biologists to browse, curate, and save the text mined PPI/PTM information. The interface should provide several key functionalities: 1) The output from multiple text mining tools should be ranked, and the display of protein network and associated text evidence should be weighted; 2) Users should be able to edit the text mining results, and the asserted knowledge should be saved in standard or convertible formats for use by different communities; 3) The interface should be simple to users with customizable options and views.

Usability testing. A major activity of iProLINK will be to facilitate interactions between text mining and user communities through annual workshops including joint workshops with existing activities, such as BioCreative and International BioCuration Meetings. An annotation workshop will allow database curators to experiment with integrating multiple text mining tools into their workflow. This will provide an opportunity for investigation of usability testing, a widely neglected topic in literature text mining. Building on the coauthors' extensive experience in evaluation of interactive systems [30], we will employ well-understood formal and informal techniques for user interface evaluation—those specific to search interfaces [31] or in general [32]—to address the lack of research into user interface design for biomedical text mining tools for curators.

5. Conclusion

We have presented a basic framework, iProLINK, to link the text mining tool developers to the ontology and systems biology user/curator communities. We used several use cases to illustrate the need and feasibility of bridging disparate communities, and analyzed requirements of the interface and major gaps in the community effort. A well designed interface and community workshops for curation and evaluation of tools will be the keys for success.

6. Acknowledgements

The work at PIR (HL, ZZH, CHW) was supported by National Science Foundation (NSF) Grant IIS-0639092, and US Army TATRC #W81XWH0720112. The work at MITRE (KBC, LH) was supported by NSF Grant II-0640153. The work of CNIO was supported by grant ENFIN NoE LSHG-CT-2005-518254.

References

[1] M.Y. Galperin. The Molecular Biology Database Collection: 2008 update. *Nucleic Acids Res.* 36(Database issue):D2-4, 2008.

[2] Z.Z. Hu, I. Mani, V. Hermoso, H. Liu and C.H. Wu. iProLINK: an integrated protein resource for literature mining. *Comput Biol Chem* 28: 409-416, 2004.

[3] C.H. Wu, L.S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, et al. The Protein Information Resource. *Nucleic Acids Research*, 31: 345-347, 2003.

[4] I. Mani, Z.Z. Hu, S.B. Jang, K. Samuel, M. Krause, J Phillips, C.H. Wu. Protein name tagging guidelines: lessons learned. *Comparative and Functional Genomics* 6:72-76. 2005.

[5] Z.Z. Hu, M. Narayanaswamy, K.E. Ravikumar, K. Vijay-Shanker, C.H. Wu. Literature mining and database annotation of protein phosphorylation using a rule-based system. *Bioinformatics* 21(11): 2759-2765, 2005.

[6] H. Liu, Z.Z. Hu, C.H. Wu. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 22:103-105, 2006.

[7] L. Hirschman, A. Yeh, C. Blaschke, and A. Valencia, Overview of BioCreative: Critical Assessment of Information Extraction for Biology, *BMC Bioinformatics* 6(Suppl 1):S1, 2005.

[8] M. Krallinger, A. Morgan, L. Smith, F. Leitner, Tanabe, et al. Evaluation of text mining systems for Biology: overview of the Second BioCreative community challenge, *Genome Biology*, 9(Suppl 2):S1, 2008.

[9] W. Hersh, A. Cohen, L. Ruslen, P. Roberts: TREC 2007 Genomics Track Overview. In *Proceedings of the Sixteenth Annual Text REtrieval Conference - TREC 2007;* 2007; Gaithersburg, MD.

[10] T.G. Dietterich: Ensemble methods in machine learning. *Multiple Classifier Systems*, 1857:1-15, 2000.

[11] Y.S., Chung, D.F. Hsu, C.Y. Tang: On the Diversity-Performance Relationship for Majority Voting in Classifier Ensembles. In: *7th International Workshop on Multiple Classifier Systems:* Springer Verlag; 2007.

[12] L. Si, T. Kanungo and X. Huang. Boosting Performance of Bio-Entity Recognition by Combining Results from Multiple Systems, In *Proc of Workshop on Data Mining in Bioinformatics*, 2005, pp. 76–83.

[13] J. Wilbur, L. Smith and L. Tanabe. BioCreative 2. Gene Mention Task, In *Proc of the Second BioCreative Challenge Evaluation Workshop*, 2007, pp. 7–16.

[14] L. Smith, L. Tanabe, R. Ando, C. Kuo, J. Chung, et al. Overview of BioCreative II gene mention recognition. *Genome Biology*, 9(Suppl 2):S2, 2008.

[15] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, J.Hakenberg, C Plake et al. Introducing meta-services for biomedical information extraction. *Genome Biology*, 9(Suppl 2):S6, 2008.

[16] GENIA Project (2001) GPML overview. http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/GPML/.

[17] K.B. Cohen, L. Fox, P.V. Ogren, and L. Hunter. Corpus design for biomedical natural language processing. Linking Biological Literature, Ontologies and Databases, 2005, pp. 38-45. Association for Computational Linguistics.

[18] D. Ferrucci and A. Lally. Building an example application with the unstructured information management architecture. *IBM Systems Journal* 43(3):455-475, 2004.

[19] R. Mack, S. Mukherjea, A. Soffer, N. Uramoto, E. Brown, et al. Text analytics for life science using the unstructured information management architecture. *IBM Systems Journal* 43(3):490-515, 2004.

[20] W.A. Baumgartner Jr., K.B. Cohen, and L. Hunter. An open-source framework for large-scale, flexible evaluation of biomedical text mining systems. *Journal of Biomedical Discovery and Collaboration* 3(1), 2008.

[21] Y. Kano, N. Nguyen, R. Sætre, K. Yoshida, et al. Filling the gaps between tools and users: A tool comparator, using protein-protein interactions as an example. *Pacific Symposium on Biocomputing,* 6:616-627, 2008.

[22] H.L. Johnson, W.A. Baumgartner Jr., M. Krallinger, K.B. Cohen, L. Hunter. Corpus refactoring: a feasibility study. *Journal of Biomedical Discovery and Collaboration* 2(4), 2006.

[23] B. Smith, M. Ashburner, C. Rosse, J Bard, W. Bug, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotechnol*. 25:1251-5, 2007.

[24] D.A. Natale, C.N. Arighi, W. Barker, J. Blake, T. Chang, et al. Framework for a Protein Ontology. *BMC Bioinformatics*, 8(Suppl 9):S1, 2007

[25] H. Huang, Z.Z. Hu, C.N. Arighi, C.H. Wu. Integration of bioinformatics resources for functional analysis of gene expression and proteomic data. *Front Biosci.* 12:5071-88, 2007.

[26] S. Kim, S.Y. Shin, I.H. Lee, S.J. Kim, R. Sriram, B.T. Zhang. PIE: an online prediction system for protein-protein interactions from text. *Nucleic Acids Res.* 36(Web Server issue):W411-5, 2008.

[27] J.M. Fernández, R. Hoffmann, A. Valencia. iHOP web services. *Nucleic Acids Res*. 35:W21-6, 2007.

[28] M.S. Cline, M. Smoot, E. Cerami, A. Kuchinsky, et al. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc.* 2:2366-82, 2007.

[29] Z.Z. Hu, H. Huang, B. Kagan, A. Riegel, A. Wellstein, A. Dritschilo, C.H. Wu. Protein-centric integration and functional analysis of cancer omics data. *US HUPO Annual Conference* March 16-19, 2008, Washington DC.

[30] J. Polifroni, L. Hirschman, S. Seneff, and V. Zue. Experiments in evaluating interactive spoken language systems. *Human Language Technology Conference: Proceedings of the Workshop on Speech and Natural Language,* 1992, pp. 28-33.

[31] M. Hearst, "Evaluation of Search Interfaces," *Modern Information Retrieval*, 2nd Edition: The Use and Technology behind Search Engines, R. Baeza-Yates, B. Ribeiro-Neto, and M. Hearst, Addison-Wesley, to appear

[32] B. Shneiderman and C. Plaisant. *Designing the User Interface.* Addison Wesley. 2004.

PAG XVII - January 10-14, 2009

http://www.intl-pag.org/

PIR (Protein Information Resource) Workshop

Sunday January 11, 8:00-10:10am

http://www.intl-pag.org/17/17-pir.html

Text Mining for Database Curation

Organizer: Cathy H. Wu, Georgetown University Medical Center
(wuc@georgetown.edu)

ABSTRACT

The biological literature represents the repository of biological knowledge. The increasing volume of scientific literature now available electronically and the exponential growth of large-scale molecular data have prompted active research in biological text mining to facilitate literature-based database curation. In particular, evidence attribution of experimentally validated information extracted from the scientific literature will become increasingly important to ensure the annotation quality of biological databases. Many text mining tools and resources have been developed. There are community efforts, such as the BioCreative Challenge Evaluations, for evaluating text mining systems applied to the biological domain. However, these tools are still not being fully utilized by the broad biological user communities. Such a gap is partly due to intrinsic complexity of biological text for mining, and partly to the lack of data standards and close interactions between the text mining and user communities to conduct utility/usability analysis and use case development. This workshop will include presentations and a panel discussion to facilitate the development of text mining systems that address the needs of the biocuration and biological research community.

SPEAKERS

8:00-8:10 am Cathy Wu
Introduction

8:10-8:40 am Carl Schmidt
Text-Mining to Aid Annotation of the Gallus Reactome

8:40-9:10 am Lynette Hirschman
BioCreative: Evaluating Text Mining for the BioCuration Workflow

9:10-9:40 am Cathy Wu
iProLINK: Linking Text Mining with Ontology and Systems Biology

9:40-10:10 am Panel discussion
Text Mining for Database Curation

Text-Mining to Aid Annotation of the Gallus Reactome

Carl J. Schmidt[1], Catalina Oana Tudor[1], Li Jin, Keith Decker[1], Peter D'Eustachio[2], and Vijay Shanker[1]

[1]University of Delaware, [2] New York University, School of Medicine
schmidtc@udel.edu, oanat@UDel.Edu, jin@mail.eecis.udel.edu, decker@cis.udel.edu, Peter.D'Eustachio@nyumc.org, vijay@cis.udel.edu

The objective of Gallus Reactome is to provide a curated set of metabolic and signaling pathways for the chicken. To assist annotators, we are developing a set of tools designed to extract and prioritize text from abstracts that are relevant to the gene products being annotated. Key terms extracted from abstracts with eGIFT are grouped by whether the key term describes the target gene product alone or describes its interaction with other proteins. The latter group is likely to be of greater importance to Reactome annotators. Since Gallus Reactome is particularly interested in papers that document pathways in the chicken, abstracts are classified according to the species that were the source of the experimental material. The annotator is provided with a web page containing sentences that have been prioritized according to the species of interest, and the likelihood that the sentences are relevant to pathways. The annotator can choose to view the complete abstract or article containing sentences that appear relevant to the current task. Sentences can also be saved to a GeneWiki page that allows the scientific community rapid access to information the annotator viewed as germane to the reaction pathway.

BioCreative: Evaluating Text Mining for the BioCuration Workflow

Lynette Hirschman
The MITRE Corporation
lynette@mitre.org

There has been increasing interest in applying text mining technology to BioCuration, but it is still difficult to point to major successful applications, or to determine what tools are available for which aspects of curation. BioCreative (Critical Assessment of Information Extraction in Biology) was organized to encourage progress in this important area through development of Challenge Evaluations, focused largely on the biocuration workflow – see the recent special issue of Genome Biology (Vol 9, Suppl 2 2008). In BioCreative II, the curators from the MINT and IntAct protein-protein interaction databases provided a "gold standard" expert curated data set against which to compare performance of text mining tools. This advanced task, developed by Martin Krallinger in Alfonso Valencia's group at CNIO, included identification of relevant articles for curation, extraction of interacting proteins (with their SwissProt identifiers), extraction of experimental methods, and identification of supporting textual evidence. In a first step towards making these tools available, many of the participants in BioCreative II have contributed their tools to a MetaServer (http://bcms.bioinfo.cnio.es/), developed by the team at the CNIO. For BioCreative III, working with Cathy Wu and the PIR curation

team, our goal is to assess text mining tools "in situ" – for example, in the context of a curation jamboree, with curators providing an evaluation of the usability and utility of the tools. BioCreative III is planned for 2009-2010, and we are soliciting input, requirements and participation from the BioCurator community.

*This work is supported by NSF Grant IIS-0844419.

iProLINK: Linking Text Mining with Ontology and Systems Biology

Cathy H. Wu
Georgetown University Medical Center
wuc@georgetown.edu

The rapid growth of scientific literature and of large-scale molecular data has prompted active research in biological text mining to facilitate literature-based database curation. Meanwhile, systems biology knowledgebases and ontologies are emerging as critical tools in biological research where complex data in disparate resources need to be integrated and annotated. PIR has developed iProLINK as a resource to support text mining research. It provides literature corpora with annotation-tagged abstracts for training and benchmarking text mining tools, as well as tools such as RLIMS-P for mining protein phosphorylation objects from MEDLINE abstracts and BioThesaurus for identification of synonymous and ambiguous gene/protein names to support named entity recognition. Built on iProLINK, PIR is developing a framework for linking text mining with ontology and systems biology, focusing on integration of public text mining tools for mining protein-protein interactions, including the post-translational modifications and pathogen-host interactions. Use cases will be presented with applications for curation of molecular and ontological data and analysis of systems biology data in the network and pathway context. The framework will facilitate the utility, usability and requirement analyses by biologists to guide future development of text mining tools and systems that will be broadly utilized by biologists for database curation and knowledge discovery.

Spectrum	UniProtKB AC	Race-P Protein	Mascot Protein
10.11_81_0001	A3NFY4	Putative uncharacterized protein	A3NFY4_BURP6 Putative uncharacterized protein (32(
10.111_90_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.111_90_0001	A3NL02	Putative uncharacterized protein	A3NL02_BURP6 Putative uncharacterized protein (32(
10.131_91_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
10.135_92_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.154_96_0001	A3NCH8	Transcriptional regulator, Sir2 family	A3NCH8_BURP6 Transcriptional regulator, Sir2 family
10.154_96_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.155_04_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.159_01_0001	A3NKJ8	Putative uncharacterized protein	A3NKJ8_BURP6 Putative uncharacterized protein (32(
10.174_03_0001	A3NH76	Putative uncharacterized protein	A3NH76_BURP6 Putative uncharacterized protein (32
10.207_80_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.25_83_0001	A3NCH8	Transcriptional regulator, Sir2 family	A3NCH8_BURP6 Transcriptional regulator, Sir2 family
10.26_84_0001	A3NIS1	Putative uncharacterized protein	A3NIS1_BURP6 Putative uncharacterized protein (320
10.27_85_0001	A3NAI6	Trigger factor	TIG_BURP6 Trigger factor (320373: Burkholderia pseu
10.27_85_0001	A3NKE1	Putative uncharacterized protein	A3NKE1_BURP6 Putative uncharacterized protein (32(
10.47_86_0001	A3NBH6	Putative uncharacterized protein	A3NBH6_BURP6 Putative uncharacterized protein (32
10.47_86_0001	A3N698	Putative uncharacterized protein	A3N698_BURP6 Putative uncharacterized protein (32(
10.62_87_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.67_88_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
10.91_89_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32(
11.108_21_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.112_22_0001	Q2SVC6	Succinyl-CoA:3-ketoacid-coenzyme A trans	Q2SVC6_BURTA 3-oxoadipate CoA-succinyl transferas
11.131_23_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.137_11_0001	Q2T2Q2	Putative uncharacterized protein	Q2T2Q2_BURTA Putative uncharacterized protein (27
11.139_06_0001	Q2SZU0	Putative phospholipase C accessory protei	Q2SZU0_BURTA Phospholipase C accessory protein, p
11.150_24_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.155_25_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.16_26_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.161_27_0001	Q2SV18	Phage integrase	Q2SV18_BURTA Phage integrase (271848: Burkholder
11.184_12_0001	Q2SU39	50S ribosomal protein L5	RL5_BURTA 50S ribosomal protein L5 (271848: Burkh
11.19_05_0001	Q2SUZ2	Type I restriction system adenine methylas	Q2SUZ2_BURTA Type I restriction system adenine me
11.191_28_0001	Q2T018	Lysozyme	Q2T018_BURTA Lysozyme (271848: Burkholderia thai
11.208_29_0001	Q2SWA7	Putative aldehyde dehydrogenase	Q2SWA7_BURTA Aldehyde dehydrogenase (271848: E
11.218_13_0001	Q2T916	Rhs1 protein	Q2T916_BURTA Rhs1 protein (271848: Burkholderia t
11.218_13_0001	Q2T711	Translocator protein bipB	BIPB_BURTA Translocator protein bipB (271848: Burk
11.22_15_0001	Q2T2D4	Putative uncharacterized protein	Q2T2D4_BURTA Putative uncharacterized protein (27
11.22_15_0001	Q2T4C3	Sensor protein	Q2T4C3_BURTA Sensor protein (271848: Burkholderia
11.240_31_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
11.241_32_0001	Q2SXF2	Phasin family protein	Q2SXF2_BURTA Phasin family protein (271848: Burkh
11.241_32_0001	Q2STI2	Site-specific recombinase, phage integrase	Q2STI2_BURTA Site-specific recombinase, phage integ
11.243_14_0001	Q2T2Q2	Putative uncharacterized protein	Q2T2Q2_BURTA Putative uncharacterized protein (27
11.47_16_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
11.49_08_0001	Q2SUZ9	TnpB protein	Q2SUZ9_BURTA TnpB protein (271848: Burkholderia t
11.57_17_0001	Q2SZU0	Putative phospholipase C accessory protei	Q2SZU0_BURTA Phospholipase C accessory protein, p
11.72_18_0001	Q2T2Q2	Putative uncharacterized protein	Q2T2Q2_BURTA Putative uncharacterized protein (27
11.78_09_0001	Q2T5X9	CmaB	Q2T5X9_BURTA CmaB (271848: Burkholderia thailanc
11.78_09_0001	Q2SZU9	Carboxymuconolactone decarboxylase fam	Q2SZU9_BURTA Carboxymuconolactone decarboxylas
11.81_19_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
11.90_10_0001	Q2T2U2	Putative uncharacterized protein	Q2T2U2_BURTA Putative uncharacterized protein (27
12.1_35_0001	A4JNK3	Putative uncharacterized protein	A4JNK3_BURVG Putative uncharacterized protein (26
12.101_45_0001	A4JE78	Efflux transporter, RND family, MFP subun	A4JE78_BURVG Efflux transporter, RND family, MFP s
12.117_46_0001	A4JLN3	Glycosyl transferase, family 2	A4JLN3_BURVG Glycosyl transferase, family 2 (26948
12.154_47_0001	A4JAR6	DNA-directed RNA polymerase subunit alp	RPOA_BURVG DNA-directed RNA polymerase subunit
12.156_48_0001	A4JNK7	Putative uncharacterized protein	A4JNK7_BURVG Putative uncharacterized protein (26
12.160_49_0001	A4JRH3	Putative uncharacterized protein	A4JRH3_BURVG Putative uncharacterized protein (26
12.167_50_0001	A4JE78	Efflux transporter, RND family, MFP subun	A4JE78_BURVG Efflux transporter, RND family, MFP s
12.171_51_0001	A4JCC6	Guanosine-3,5-bis(diphosphate) 3-pyrophc	A4JCC6_BURVG (P)ppGpp synthetase I, SpoT/RelA (26

12.197_52_0001	A4JNK7	Putative uncharacterized protein	A4JNK7_BURVG Putative uncharacterized protein (26
12.198_57_0001	A4JD54	Putative uncharacterized protein	A4JD54_BURVG Putative uncharacterized protein (26
12.198_57_0001	A4JFW6	Phage integrase family protein	A4JFW6_BURVG Phage integrase family protein (2694
12.218_53_0001	A4JT92	RNA-directed DNA polymerase (Reverse tr	A4JT92_BURVG RNA-directed DNA polymerase (Reve
12.22_36_0001	A4JRC7	Putative uncharacterized protein	A4JRC7_BURVG Putative uncharacterized protein (26
12.22_36_0001	A4JT51	Prolyl aminopeptidase	A4JT51_BURVG Prolyl aminopeptidase (269482: Burkl
12.26_33_0001	A4JQW6	Putative uncharacterized protein	A4JQW6_BURVG Putative uncharacterized protein (26
12.28_37_0001	A4JNK3	Putative uncharacterized protein	A4JNK3_BURVG Putative uncharacterized protein (26
12.39_38_0001	A4JSF9	Putative uncharacterized protein	A4JSF9_BURVG Putative uncharacterized protein (269
12.39_38_0001	A4JW42	Putative uncharacterized protein	A4JW42_BURVG Putative uncharacterized protein (26
12.46_39_0001	A4JDC4	Putative uncharacterized protein	A4JDC4_BURVG Putative uncharacterized protein (26
12.73_40_0001	A4JE78	Efflux transporter, RND family, MFP subun	A4JE78_BURVG Efflux transporter, RND family, MFP s
12.83_42_0001	A4JT92	RNA-directed DNA polymerase (Reverse tr	A4JT92_BURVG RNA-directed DNA polymerase (Reve
12.83_42_0001	A4JPC6	GP32 family protein	A4JPC6_BURVG GP32 family protein (269482: Burkho
12.89_43_0001	A4JT92	RNA-directed DNA polymerase (Reverse tr	A4JT92_BURVG RNA-directed DNA polymerase (Reve
12.9_34_0001	A4JE78	Efflux transporter, RND family, MFP subun	A4JE78_BURVG Efflux transporter, RND family, MFP s
12.93_54_0001	A4JE78	Efflux transporter, RND family, MFP subun	A4JE78_BURVG Efflux transporter, RND family, MFP s
12.99_44_0001	A4JPR6	Putative uncharacterized protein	A4JPR6_BURVG Putative uncharacterized protein (26
6.102_04_0001	A4JQL9	Putative uncharacterized protein	A4JQL9_BURVG Putative uncharacterized protein (26
6.118_05_0001	A4JTG8	Putative uncharacterized protein	A4JTG8_BURVG Putative uncharacterized protein (26
6.118_05_0001	A4JV87	DNA topoisomerase	A4JV87_BURVG DNA topoisomerase (269482: Burkho
6.125_06_0001	A4JPQ7	Response regulator receiver protein	A4JPQ7_BURVG Response regulator receiver protein (
6.136_07_0001	A4JW67	Putative uncharacterized protein	A4JW67_BURVG Putative uncharacterized protein (26
6.15_08_0001	A4JFW6	Phage integrase family protein	A4JFW6_BURVG Phage integrase family protein (2694
6.69_01_0001	A4JPQ7	Response regulator receiver protein	A4JPQ7_BURVG Response regulator receiver protein (
6.69_01_0002	A4JNV7	Putative transposase	A4JNV7_BURVG Putative transposase (269482: Burkh
6.99_03_0001	A4JAC2	Transposase, Helix-turn-helix, type 11 dom	A4JAC2_BURVG Helix-turn-helix, type 11 domain prot
7.101_14_0001	Q62KI2	Threonyl-tRNA synthetase	SYT_BURMA Threonyl-tRNA synthetase (13373: Burkh
7.132_15_0001	Q62C20	Putative uncharacterized protein	Q62C20_BURMA Putative uncharacterized protein (13
7.134_21_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
7.142_22_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
7.153_23_0001	Q4V2Q0	Hcp1	Q4V2Q0_BURMA Hcp1 (13373: Burkholderia mallei)
7.153_23_0001	Q4V296	Sigma-54 dependent transcriptional regula	Q4V296_BURMA Sigma-54 dependent transcriptional
7.153_23_0001	Q62GL9	30S ribosomal protein S8	RS8_BURMA 30S ribosomal protein S8 (13373: Burkho
7.167_16_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
7.167b_17_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
7.167b_17_0001	Q62IX8	Pyruvate dehydrogenase, E1 component	Q62IX8_BURMA Pyruvate dehydrogenase, E1 compor
7.167b_17_0001	Q62K96	Putative uncharacterized protein	Q62K96_BURMA Putative uncharacterized protein (13
7.178_38_0001	Q62GG1	Putative uncharacterized protein	Q62GG1_BURMA Putative uncharacterized protein (1
7.179_39_0001	Q62CQ2	Sensor protein	Q62CQ2_BURMA Sensor protein (13373: Burkholderia
7.19_28_0001	Q62AV7	cellulose synthase operon protein C	Q62AV7_BURMA Cellulose synthase operon protein C
7.20_29_0001	Q62LX9	Isocitrate dehydrogenase [NADP]	Q62LX9_BURMA Isocitrate dehydrogenase [NADP] (13
7.20_29_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
7.27_30_0001	Q62K99	Putative Syringomycin biosynthesis enzym	Q62K99_BURMA Syringomycin biosynthesis enzyme,
7.31_10_0001	Q62JC5	Elongation factor Ts	EFTS_BURMA Elongation factor Ts (13373: Burkholder
7.31_10_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
7.59_11_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
7.63_33_0001	Q4V2Q0	Hcp1	Q4V2Q0_BURMA Hcp1 (13373: Burkholderia mallei)
7.69_34_0001	Q62KK6	Pseudouridine synthase	Q62KK6_BURMA Pseudouridine synthase (13373: Bur
7.9_26_0001	Q62KA8	L-ornithine 5-monooxygenase	Q62KA8_BURMA L-ornithine 5-monooxygenase (1337
7.95_12_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
7.97_13_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
7.97_13_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
8.12_57_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
8.124_58_0001	Q62G22	Adenosylhomocysteinase	SAHH_BURMA Adenosylhomocysteinase (13373: Burk
8.131_49_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
8.131_49_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
8.133_50_0001	Q62LZ3	Aminopeptidase N	Q62LZ3_BURMA Aminopeptidase N (13373: Burkhold

8.133_50_0001	Q62EG1	H-NS histone family protein	Q62EG1_BURMA H-NS histone family protein (13373:
8.136_51_0001	Q62EQ4	Argininosuccinate synthase	ASSY_BURMA Argininosuccinate synthase (13373: Bu
8.136_51_0001	Q62B16	Effector protein bopA	BOPA_BURMA Effector protein bopA (13373: Burkhol
8.141_40_0001	Q629S0	Putative uncharacterized protein	Q629S0_BURMA Putative uncharacterized protein (13
8.144_41_0001	Q62KW4	DNA helicase II	Q62KW4_BURMA DNA helicase II (13373: Burkholder
8.16_45_0001	Q62BP9	Putative uncharacterized protein	Q62BP9_BURMA Putative uncharacterized protein (13
8.161_52_0001	Q62IY3	Phasin family protein	Q62IY3_BURMA Phasin family protein (13373: Burkhc
8.161_52_0001	Q62GL1	30S ribosomal protein S3	RS3_BURMA 30S ribosomal protein S3 (13373: Burkhc
8.167_59_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
8.167_59_0001	Q62ED9	Putative uncharacterized protein	Q62ED9_BURMA Putative uncharacterized protein (13
8.172_53_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
8.181_60_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
8.183_42_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
8.191_43_0001	Q62CT8	Putative uncharacterized protein	Q62CT8_BURMA Putative uncharacterized protein (13
8.191_43_0001	Q62GK8	50S ribosomal protein L2	RL2_BURMA 50S ribosomal protein L2 (13373: Burkhc
8.206_44_0001	Q62GL8	30S ribosomal protein S14	RS14_BURMA 30S ribosomal protein S14 (13373: Burl
8.29_55_0001	Q62ED9	Putative uncharacterized protein	Q62ED9_BURMA Putative uncharacterized protein (13
8.30_46_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
8.318_54_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
8.53_47_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
8.53_47_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
8.57_56_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
8.78_48_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
9.1_74_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
9.119_70_0001	A2RZ02	Putative uncharacterized protein	A2RZ02_BURM9 Putative uncharacterized protein (41
9.12_68_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
9.12_71_0001	A2RXT6	Putative uncharacterized protein	A2RXT6_BURM9 Putative uncharacterized protein (41
9.121_72_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
9.121_72_0001	A2RWN4	Polyphosphate kinase 2	A2RWN4_BURM9 Polyphosphate kinase 2 (412022: B
9.162_63_0001	A2SBG2	Trigger factor	TIG_BURM9 Trigger factor (412022: Burkholderia mal
9.162_63_0001	A2S1N6	Type III secretion system transcriptional re	A2S1N6_BURM9 Type III secretion system transcriptic
9.164_73_0001	A2S472	Ketol-acid reductoisomerase	ILVC_BURM9 Ketol-acid reductoisomerase (412022: B
9.165_76_0001	A2S4H3	Transaldolase	A2S4H3_BURM9 Transaldolase (412022: Burkholderia
9.165_76_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
9.165_76_0001	A2S0S0	Putative uncharacterized protein	A2S0S0_BURM9 Putative uncharacterized protein (41
9.182_77_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
9.183_64_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
9.279_65_0001	A2RXT6	Putative uncharacterized protein	A2RXT6_BURM9 Putative uncharacterized protein (41
9.294_79_0001	A2RZ02	Putative uncharacterized protein	A2RZ02_BURM9 Putative uncharacterized protein (41
9.294_79_0001	A2S3Z5	Putative uncharacterized protein	A2S3Z5_BURM9 Putative uncharacterized protein (41
9.3_67_0001	A2RZL5	GMC oxidoreductase	A2RZL5_BURM9 Putative cholesterol oxidase (412022
9.48_75_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
9.78_69_0001	A2SAM0	D-alanyl-D-alanine carboxypeptidase famil	A2SAM0_BURM9 D-alanyl-D-alanine carboxypeptidas
Usamrid__01_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__01_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
Usamrid__01_0001	Q62KY2	Branched-chain amino acid ABC transporte	Q62KY2_BURMA Branched-chain amino acid ABC trar
Usamrid__01_0001	Q62KP7	Putative uncharacterized protein	Q62KP7_BURMA Putative uncharacterized protein (13
Usamrid__01_0001	Q62JX9	Putative uncharacterized protein	Q62JX9_BURMA Putative uncharacterized protein (13
Usamrid__05_0001	Q62KD9	Arginine deiminase	ARCA_BURMA Arginine deiminase (13373: Burkholde
Usamrid__05_0001	Q62JD3	Outer membrane protein, OmpH/HlpA fan	Q62JD3_BURMA Outer membrane protein, OmpH/Hl
Usamrid__05_0001	Q62GK8	50S ribosomal protein L3	RL2_BURMA 50S ribosomal protein L2 (13373: Burkhc
Usamrid__06_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__06_0001	Q62JD3	Outer membrane protein, OmpH/HlpA fan	Q62JD3_BURMA Outer membrane protein, OmpH/Hl
Usamrid__07_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__07_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
Usamrid__07_0001	Q62JX9	Putative uncharacterized protein	Q62JX9_BURMA Putative uncharacterized protein (13

Usamrid__08_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__08_0001	Q4V2D7	Putative uncharacterized protein	Q4V2D7_BURMA Putative uncharacterized protein (1
Usamrid__08_0001	Q62JC7	Ribosome-recycling factor	RRF_BURMA Ribosome-recycling factor (13373: Burkh
Usamrid__08_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__09_0001	Q62JK6	Trigger factor	TIG_BURMA Trigger factor (13373: Burkholderia malle
Usamrid__09_0001	Q62C71	Putative uncharacterized protein	Q62C71_BURMA Putative uncharacterized protein (13
Usamrid__09_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__09_0001	Q62CK6	UvrABC system protein B	UVRB_BURMA UvrABC system protein B (13373: Burk
Usamrid__10_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__10_0001	Q62B07	Translocator protein bipB	BIPB_BURMA Translocator protein bipB (13373: Burkh
Usamrid__11_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__11_0001	Q62AV7	cellulose synthase operon protein C	Q62AV7_BURMA Cellulose synthase operon protein C
Usamrid__11_0001	Q62CR4	Isovaleryl-CoA dehydrogenase	Q62CR4_BURMA Isovaleryl-CoA dehydrogenase (1337
Usamrid__13_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__13_0001	Q62B07	Translocator protein bipB	BIPB_BURMA Translocator protein bipB (13373: Burkh
Usamrid__13_0001	Q4V2C6	Putative uncharacterized protein	Q4V2C6_BURMA Putative uncharacterized protein (13
Usamrid__13_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
Usamrid__13_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
Usamrid__14_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__14_0001	Q4V2S5	Putative uncharacterized protein	Q4V2S5_BURMA Putative uncharacterized protein (13
Usamrid__14_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__14_0001	Q62JL5	Putative uncharacterized protein	Q62JL5_BURMA Putative uncharacterized protein (13
Usamrid__15_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
Usamrid__16_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__18_0001	Q62EW4	Ferredoxin--NADP reductase	Q62EW4_BURMA Ferredoxin--NADP reductase (1337
Usamrid__18_0001	Q62M00	Putative uncharacterized protein	Q62M00_BURMA Putative uncharacterized protein (1
Usamrid__19_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__19_0001	Q62GK8	50S ribosomal protein L4	RL2_BURMA 50S ribosomal protein L2 (13373: Burkho
Usamrid__21_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__22_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__22_0001	Q62IN1	Polyribonucleotide nucleotidyltransferase	PNP_BURMA Polyribonucleotide nucleotidyltransfera
Usamrid__23_0001	Q62LV0	Pseudouridine synthase	Q62LV0_BURMA Pseudouridine synthase (13373: Bur
Usamrid__24_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__25_0001	Q62I82	60 kDa chaperonin	Q4PPC2_BURMA 60 kDa chaperonin (13373: Burkhold
Usamrid__25_0001	Q62I82	60 kDa chaperonin	CH60_BURMA 60 kDa chaperonin (13373: Burkholder
Usamrid__25_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__26_0001	Q62FN4	Putative uncharacterized protein	Q62FN4_BURMA Putative uncharacterized protein (13
Usamrid__27_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__27_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
Usamrid__28_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__28_0001	Q62D46	Putative uncharacterized protein	Q62D46_BURMA Putative uncharacterized protein (13
Usamrid__28_0001	Q62LX9	Isocitrate dehydrogenase [NADP]	Q62LX9_BURMA Isocitrate dehydrogenase [NADP] (13
Usamrid__29_0001	Q62HP5	Putative uncharacterized protein	Q62HP5_BURMA Putative uncharacterized protein (13
Usamrid__29_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (13
Usamrid__29_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (13
Usamrid__29_0001	Q62H93	Putative uncharacterized protein	Q62H93_BURMA Putative uncharacterized protein (13
Usamrid__29_0001	Q62K96	Putative uncharacterized protein	Q62K96_BURMA Putative uncharacterized protein (13
Usamrid__29_0001	Q62K99	Putative Syringomycin biosynthesis enzym	Q62K99_BURMA Syringomycin biosynthesis enzyme,
Usamrid__29_0001	Q4V2D7	Putative uncharacterized protein	Q4V2D7_BURMA Putative uncharacterized protein (1
Usamrid__29_0001	Q4V2B6	Putative uncharacterized protein	Q4V2B6_BURMA Putative uncharacterized protein (13
Usamrid__31_0001	Q62BP9	Putative uncharacterized protein	Q62BP9_BURMA Putative uncharacterized protein (13

Usamrid__33_0001	Q62I76	Putative uncharacterized protein	Q62I76_BURMA Putative uncharacterized protein (13
Usamrid__34_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__34_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__34_0001	Q4V2G9	Putative uncharacterized protein	Q4V2G9_BURMA Putative uncharacterized protein (1
Usamrid__34_0001	Q4V2B6	Putative uncharacterized protein	Q4V2B6_BURMA Putative uncharacterized protein (1
Usamrid__35_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__35_0001	Q62ED9	Putative uncharacterized protein	Q62ED9_BURMA Putative uncharacterized protein (1
Usamrid__35_0001	Q62JX9	Putative uncharacterized protein	Q62JX9_BURMA Putative uncharacterized protein (13
Usamrid__35_0001	Q62IH9	Putative uncharacterized protein	Q62IH9_BURMA Putative uncharacterized protein (13
Usamrid__36_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__37_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__38_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__39_0001	Q4V2B6	Putative uncharacterized protein	Q4V2B6_BURMA Putative uncharacterized protein (1
Usamrid__39_0001	Q62A78	Putative uncharacterized protein	Q62A78_BURMA Putative uncharacterized protein (1
Usamrid__40_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__40_0001	Q62HI5	Exodeoxyribonuclease 7 large subunit	Q62HI5_BURMA Exodeoxyribonuclease 7 large subun
Usamrid__40_0001	Q62JX9	Putative uncharacterized protein	Q62JX9_BURMA Putative uncharacterized protein (13
Usamrid__41_0001	Q62LV0	Pseudouridine synthase	Q62LV0_BURMA Pseudouridine synthase (13373: Bur
Usamrid__42_0001	Q4V2B6	Putative uncharacterized protein	Q4V2B6_BURMA Putative uncharacterized protein (1
Usamrid__42_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__42_0001	Q62K96	Putative uncharacterized protein	Q62K96_BURMA Putative uncharacterized protein (1
Usamrid__43_0001	Q62B16	Effector protein bopA	BOPA_BURMA Effector protein bopA (13373: Burkhol
Usamrid__43_0001	Q62HM5	50S ribosomal protein L28	RL28_BURMA 50S ribosomal protein L28 (13373: Burk
Usamrid__43_0001	Q62G91	Putative uncharacterized protein	Q62G91_BURMA Putative uncharacterized protein (1
Usamrid__43_0001	Q62J71	Molybdenum cofactor biosynthesis protei	Q62J71_BURMA Molybdenum cofactor biosynthesis p
Usamrid__44_0001	Q62A00	Putative uncharacterized protein	Q62A00_BURMA Putative uncharacterized protein (1
Usamrid__44_0001	Q62BP9	Putative uncharacterized protein	Q62BP9_BURMA Putative uncharacterized protein (1
Usamrid__45_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
Usamrid__45_0001	A2S6F9	Putative uncharacterized protein	A2S6F9_BURM9 Putative uncharacterized protein (41
Usamrid__45_0001	A2RZI8	Conserved domain protein	A2RZI8_BURM9 Conserved domain protein (412022: I
Usamrid__45_0001	A2S2S4	Putative uncharacterized protein	A2S2S4_BURM9 Putative uncharacterized protein (41
Usamrid__47_0001	A2SAG4	Putative uncharacterized protein	A2SAG4_BURM9 Putative uncharacterized protein (41
Usamrid__48_0001	A2RXP7	Probable acyl-coenzyme A carboxylase, bic	A2RXP7_BURM9 Putative acetyl-CoA carboxylase, bio
Usamrid__48_0001	A2S511	Exodeoxyribonuclease 7 large subunit	A2S511_BURM9 Exodeoxyribonuclease 7 large subun
Usamrid__48_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
Usamrid__51_0001	A2RZ02	Putative uncharacterized protein	A2RZ02_BURM9 Putative uncharacterized protein (41
Usamrid__54_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
Usamrid__55_0001	A2S4Y3	Isocitrate dehydrogenase [NADP]	A2S4Y3_BURM9 Isocitrate dehydrogenase [NADP] (41
Usamrid__55_0001	A2S2I2	Ribonuclease R	A2S2I2_BURM9 Ribonuclease R (412022: Burkholderia
Usamrid__56_0001	A2RZC3	Putative uncharacterized protein	A2RZC3_BURM9 Putative uncharacterized protein (41
Usamrid__56_0001	A2RX76	Putative uncharacterized protein	A2RX76_BURM9 Putative uncharacterized protein (41
Usamrid__57_0001	A2SAM0	D-alanyl-D-alanine carboxypeptidase famil	A2SAM0_BURM9 D-alanyl-D-alanine carboxypeptidas
Usamrid__57_0001	A2S4J8	Putative uncharacterized protein	A2S4J8_BURM9 Putative uncharacterized protein (41
Usamrid__57_0001	A2S9G6	Putative uncharacterized protein	A2S9G6_BURM9 Putative uncharacterized protein (41
Usamrid__58_0001	A2S967	Putative uncharacterized protein	A2S967_BURM9 Putative uncharacterized protein (41
Usamrid__59_0001	A3NAH7	Putative uncharacterized protein	A3NAH7_BURP6 Putative uncharacterized protein (32
Usamrid__63_0001	A3NAV1	Putative uncharacterized protein	A3NAV1_BURP6 RNA pseudouridine synthase family p
Usamrid__63_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (32
Usamrid__63_0001	A3NAF4	Putative uncharacterized protein	A3NAF4_BURP6 Putative uncharacterized protein (32
Usamrid__63_0001	A3NAC6	Putative uncharacterized protein	A3NAC6_BURP6 Putative uncharacterized protein (32
Usamrid__63_0001	A3NMH3	Zinc-containing alcohol dehydrogenase su	A3NMH3_BURP6 Zinc-containing alcohol dehydrogen
Usamrid__63_0001	A3NGM7	Putative uncharacterized protein	A3NGM7_BURP6 Putative uncharacterized protein (3
Usamrid__63_0001	A3NFE4	Putative uncharacterized protein	A3NFE4_BURP6 Putative uncharacterized protein (32
Usamrid__64_0001	A3N7W9	Putative uncharacterized protein	A3N7W9_BURP6 Pentapeptide mxkdx repeat protein

Usamrid__65_0001	A3NAU5	Ribosome-recycling factor	RRF_BURP6 Ribosome-recycling factor (320373: Burkl
Usamrid__66_0001	A3NEG6	30S ribosomal protein S14	RS14_BURP6 30S ribosomal protein S14 (320373: Bur
Usamrid__67_0001	A3NIS1	Putative uncharacterized protein	A3NIS1_BURP6 Putative uncharacterized protein (320
Usamrid__67_0001	A3NFY4	Putative uncharacterized protein	A3NFY4_BURP6 Putative uncharacterized protein (320
Usamrid__67_0001	A3NAH7	Putative uncharacterized protein	A3NAH7_BURP6 Putative uncharacterized protein (32
Usamrid__68_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (320
Usamrid__68_0001	A3NFE4	Putative uncharacterized protein	A3NFE4_BURP6 Putative uncharacterized protein (320
Usamrid__68_0001	A3N920	Transcriptional regulator, GntR family	A3N920_BURP6 Transcriptional regulator, GntR family
Usamrid__69_0001	A3NAH7	Putative uncharacterized protein	A3NAH7_BURP6 Putative uncharacterized protein (32
Usamrid__69_0001	A3NL65	Putative uncharacterized protein	A3NL65_BURP6 Putative uncharacterized protein (320
Usamrid__70_0001	A3NLR8	Putative uncharacterized protein	A3NLR8_BURP6 Putative uncharacterized protein (320
Usamrid__70_0001	A3NI21	Putative uncharacterized protein	A3NI21_BURP6 Putative uncharacterized protein (320
Usamrid__71_0001	A3NED6	Putative PilN protein	A3NED6_BURP6 Putative PilN protein (320373: Burkh
Usamrid__72_0001	A3NH76	Putative uncharacterized protein	A3NH76_BURP6 Putative uncharacterized protein (32
Usamrid__72_0001	A3NB06	Molybdenum cofactor biosynthesis protein	A3NB06_BURP6 Molybdenum cofactor biosynthesis p
Usamrid__72_0001	A3NAH7	Putative uncharacterized protein	A3NAH7_BURP6 Putative uncharacterized protein (32
Usamrid__72_0001	A3NCH6	Putative uncharacterized protein	A3NCH6_BURP6 Putative uncharacterized protein (32
Usamrid__72_0001	A3N749	Putative uncharacterized protein	A3N749_BURP6 Putative uncharacterized protein (320
Usamrid__76_0001	Q2T205	Multifunctional CCA protein	Q2T205_BURTA tRNA nucleotidyltransferase (271848
Usamrid__76_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
Usamrid__77_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
Usamrid__78_0001	Q2T711	Translocator protein bipB	BIPB_BURTA Translocator protein bipB (271848: Burk
Usamrid__79_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
Usamrid__79_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
Usamrid__79_0001	Q2T8E3	IS407A, transposase OrfA	Q2T8E3_BURTA IS407A, transposase OrfA (271848: Bu
Usamrid__79_0001	Q2SV18	Phage integrase	Q2SV18_BURTA Phage integrase (271848: Burkholder
Usamrid__80_0001	Q2SWZ7	Elongation factor Ts	EFTS_BURTA Elongation factor Ts (271848: Burkholde
Usamrid__82_0001	Q2T1R7	Sensor histidine kinase	Q2T1R7_BURTA Sensor histidine kinase (271848: Burk
Usamrid__83_0001	Q2T4T2	Putative uncharacterized protein	Q2T4T2_BURTA Putative uncharacterized protein (27
Usamrid__84_0001	Q2T916	Rhs1 protein	Q2T916_BURTA Rhs1 protein (271848: Burkholderia t
Usamrid__85_0001	Q2T7B5	Putative Syringomycin biosynthesis enzym	Q2T7B5_BURTA Syringomycin biosynthesis enzyme, p
Usamrid__87_0001	Q2T916	Rhs1 protein	Q2T916_BURTA Rhs1 protein (271848: Burkholderia t
Usamrid__87_0001	Q2SYC2	Putative uncharacterized protein	Q2SYC2_BURTA Putative uncharacterized protein (27
Usamrid__88_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
Usamrid__90_0001	Q2T703	Effector protein bopA	BOPA_BURTA Effector protein bopA (271848: Burkho
Usamrid__90_0001	Q2T1W2	Argininosuccinate synthase	ASSY_BURTA Argininosuccinate synthase (271848: Bu
Usamrid__92_0001	A4JAP6	30S ribosomal protein S3	RS3_BURVG 30S ribosomal protein S3 (269482: Burkh
Usamrid__93_0001	A4JW42	Putative uncharacterized protein	A4JW42_BURVG Putative uncharacterized protein (26
Usamrid__93_0001	A4JF74	Elongation factor Ts	EFTS_BURVG Elongation factor Ts (269482: Burkholde
Usamrid__93_0001	A4JMJ1	Phage putative head morphogenesis prote	A4JMJ1_BURVG Phage putative head morphogenesis
Usamrid__94_0001	A4JPC6	GP32 family protein	A4JPC6_BURVG GP32 family protein (269482: Burkho
Usamrid__95_0001	A4JG61	Phage integrase domain protein	A4JG61_BURVG Phage integrase domain protein SAM
Usamrid__96_0001	A4JD20	Putative uncharacterized protein	A4JD20_BURVG Putative uncharacterized protein (26
Usamrid__96_0001	A4JGC8	NADH-quinone oxidoreductase subunit C	NUOC_BURVG NADH-quinone oxidoreductase subuni

Score	Threshold	Expect	PeptideMatche	Organism used in experiment	Identifed Strains	TaxID	Un/Induced
59	51	0.0089	56	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
90	51	0.0000066	124	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
55	51	0.023	59	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
43	50	0.31	133	B. thailandensis E264	B. thailandensis E264	271848	Induced
63	51	0.0033	129	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
53	51	0.035	47	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
53	51	0.038	109	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
64	51	0.003	103	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
49	51	0.083	45	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
49	51	0.1	79	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
60	51	0.0077	128	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
50	51	0.066	48	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
54	51	0.032	66	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
63	51	0.0041	128	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
54	51	0.03	137	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
58	51	0.011	69	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
53	51	0.036	82	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
45	51	0.26	121	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
57	51	0.013	127	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
65	51	0.0021	133	B. pseudomallei 1126B	B. pseudomallei 668	320373	Induced
52	50	0.036	118	B. thailandensis E264	B. thailandensis E264	271848	Induced
46	50	0.13	70	B. thailandensis E264	B. thailandensis E264	271848	Induced
51	50	0.042	126	B. thailandensis E254	B. thailandensis E264	271848	Induced
46	50	0.14	64	B. thailandensis E254	B. thailandensis E264	271848	Inuduced
45	50	0.17	78	B. thailandensis E254	B. thailandensis E264	271848	Induced
55	50	0.016	152	B. thailandensis E264	B. thailandensis E264	271848	Induced
54	50	0.022	149	B. thailandensis E264	B. thailandensis E264	271848	Induced
55	50	0.019	136	B. thailandensis E264	B. thailandensis E264	271848	Induced
45	50	0.19	139	B. thailandensis E264	B. thailandensis E264	271848	Induced
49	50	0.077	60	B. thailandensis E264	B. thailandensis E264	271848	Induced
62	50	0.0038	139	B. thailandensis E264	B. thailandensis E264	271848	Induced
44	50	0.21	56	B. thailandensis E264	B. thailandensis E264	271848	Induced
53	50	0.027	96	B. thailandensis E264	B. thailandensis E264	271848	Induced
60	50	0.0062	210	B. thailandensis E264	B. thailandensis E264	271848	Induced
53	50	0.027	186	B. thailandensis E264	B. thailandensis E264	271848	Induced
56	50	0.015	45	B. thailandensis E264	B. thailandensis E264	271848	Induced
54	50	0.021	130	B. thailandensis E264	B. thailandensis E264	271848	Induced
42	50	0.33	166	B. thailandensis E264	B. thailandensis E264	271848	Induced
80	50	0.00006	47	B. thailandensis E264	B. thailandensis E264	271848	Induced
54	50	0.023	48	B. thailandensis E264	B. thailandensis E264	271848	Induced
58	50	0.0094	75	B. thailandensis E264	B. thailandensis E264	271848	Induced
63	50	0.0031	174	B. thailandensis E264	B. thailandensis E264	271848	Induced
41	50	0.43	45	B. thailandensis E264	B. thailandensis E264	271848	Induced
53	50	0.028	83	B. thailandensis E264	B. thailandensis E264	271848	Induced
59	50	0.0068	80	B. thailandensis E264	B. thailandensis E264	271848	Induced
71	50	0.00041	74	B. thailandensis E264	B. thailandensis E264	271848	Induced
54	50	0.025	91	B. thailandensis E264	B. thailandensis E264	271848	Induced
55	50	0.019	130	B. thailandensis E264	B. thailandensis E264	271848	Induced
45	50	0.17	34	B. thailandensis E264	B. thailandensis E264	271848	Induced
41	51	0.58	55	B. thailandensis E264	B. thailandensis E264	271848	Induced
41	51	0.63	126	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
44	51	0.3	81	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
66	51	0.0019	90	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
46	51	0.18	89	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
39	51	0.91	32	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
45	51	0.22	115	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
51	51	0.065	192	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced

74	51	0.00031	75	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
51	51	0.055	194	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
41	51	0.55	137	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
67	51	0.0016	172	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
52	51	0.05	88	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
39	51	0.93	92	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
43	51	0.36	48	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
51	51	0.065	55	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
67	51	0.0014	92	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
57	51	0.015	136	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
51	51	0.066	125	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
42	51	0.5	118	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
57	51	0.014	171	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
53	51	0.037	102	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
55	51	0.024	160	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
53	51	0.034	126	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
49	51	0.095	116	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
47	51	0.13	113	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Induced
49	51	0.098	63	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
52	51	0.05	60	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
41	51	0.6	188	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
50	51	0.079	145	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
45	51	0.22	143	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
55	51	0.026	199	B. vietnamensis FCO369	B. vietnamiensis G4	269482	NR
57	51	0.014	108	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
44	51	0.32	88	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
53	51	0.04	141	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
50	49	0.045	136	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
47	49	0.11	58	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
46	49	0.13	83	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
41	49	0.37	65	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
77	49	0.000088	85	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
54	49	0.02	110	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
50	49	0.049	66	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
51	49	0.038	93	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
61	49	0.0035	256	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
54	49	0.018	203	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
52	49	0.028	126	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
46	49	0.12	38	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
45	49	0.17	157	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
48	49	0.084	266	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
60	49	0.0053	120	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
55	49	0.014	75	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
62	49	0.0034	99	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
76	49	0.00013	90	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
51	49	0.043	71	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
55	49	0.014	84	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
62	49	0.0028	64	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
48	49	0.086	121	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
44	49	0.19	116	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
48	49	0.075	83	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
66	49	0.0013	76	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
55	49	0.015	90	B. mallei GB8	B. mallei ATCC 23344	243160	Induced
46	49	0.14	72	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
73	49	0.00027	112	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
53	49	0.025	253	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
53	49	0.025	158	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
57	49	0.0096	155	B. mallei GB6	B. mallei ATCC 23344	243160	Induced

52	49	0.03	49	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
133	49	2.4E-10	150	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
58	49	0.0078	193	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
51	49	0.038	70	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
56	49	0.013	150	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
58	49	0.0084	180	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
55	49	0.016	56	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
51	49	0.039	105	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
54	49	0.021	264	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
50	49	0.048	93	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
48	49	0.078	95	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
47	49	0.11	82	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
43	49	0.23	99	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
55	49	0.016	63	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
54	49	0.018	94	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
47	49	0.11	54	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
49	49	0.068	74	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
60	49	0.0047	154	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
56	49	0.013	244	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
69	49	0.00062	256	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
55	49	0.014	175	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
57	49	0.01	64	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
57	49	0.0094	256	B. mallei GB6	B. mallei ATCC 23344	243160	Induced
51	50	0.042	68	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
42	50	0.35	63	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
59	50	0.0061	75	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
66	50	0.0013	104	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
50	50	0.053	62	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
44	50	0.23	91	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
103	50	0.00000027	143	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
56	50	0.013	67	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
66	50	0.0013	102	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
65	50	0.0018	121	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
55	50	0.018	85	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
53	50	0.029	147	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
66	50	0.0012	83	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
53	50	0.029	82	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
58	50	0.0079	111	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
57	50	0.0097	76	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
51	50	0.044	100	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
44	50	0.2	161	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
47	50	0.097	72	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
39	50	0.68	137	B. mallei GB5	B. mallei NCTC 10229	412022	Induced
61	49	0.0043	301	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
60	49	0.0053	190	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.033	118	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.04	108	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
50	49	0.048	75	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
60	49	0.0047	123	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
53	49	0.023	106	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.035	132	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
88	49	0.0000086	341	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.037	113	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
67	49	0.00088	335	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
58	49	0.0073	208	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
50	49	0.047	83	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced

62	49	0.0032	141	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.017	201	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
53	49	0.025	114	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.034	324	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
114	49	0.000000019	181	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
61	49	0.0041	153	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
57	49	0.0088	300	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
54	49	0.018	247	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
70	49	0.00048	297	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
53	49	0.023	225	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
66	49	0.0012	285	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.015	327	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.017	107	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
61	49	0.0043	304	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
53	49	0.025	243	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.034	57	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.04	200	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.042	111	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
60	49	0.0052	325	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
56	49	0.013	163	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.015	137	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
50	49	0.045	63	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.036	105	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
66	49	0.0012	297	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
91	49	0.0000043	98	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.031	91	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
70	49	0.00048	234	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.014	116	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
54	49	0.02	123	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
53	49	0.022	134	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
50	49	0.046	191	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
56	49	0.012	139	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
68	49	0.00071	327	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
78	49	0.000086	187	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
78	49	0.000086	187	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
58	49	0.0077	314	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.029	164	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
67	49	0.00099	135	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
63	49	0.0025	187	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
58	49	0.0075	259	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
57	49	0.0088	120	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.037	140	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
66	49	0.0013	107	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
60	49	0.0049	200	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
59	49	0.0055	294	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
57	49	0.01	62	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
55	49	0.015	152	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
54	49	0.018	99	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.036	163	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
51	49	0.042	128	B. mallei GB8	B. mallei ATCC 23344	243160	Uninduced
52	49	0.028	181		B. mallei ATCC 23344	243160	Uninduced

57	49	0.0092	101	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
91	49	0.0000038	303	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
57	49	0.011	118	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
54	49	0.019	115	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
50	49	0.048	128	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
95	49	0.0000014	336	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
66	49	0.0012	101	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
56	49	0.014	88	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
54	49	0.02	139	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
58	49	0.0084	301	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
64	49	0.002	294	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
75	49	0.00017	336	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
64	49	0.0022	137	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
57	49	0.0092	200	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
77	49	0.0001	325	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
59	49	0.0061	176	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
51	49	0.037	78	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
44	49	0.18	144	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
61	49	0.004	132	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
55	49	0.015	302	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
52	49	0.03	151	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
72	49	0.0003	228	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
57	49	0.0088	43	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
54	49	0.02	69	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
52	49	0.03	227	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
68	49	0.0007	320	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
59	49	0.0064	212	B. mallei GB6	B. mallei ATCC 23344	243160	Uninduced
52	50	0.033	115	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
52	50	0.036	156	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.039	269	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.045	151	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.039	129	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
61	50	0.0045	152	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
53	50	0.028	142	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.045	77	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
66	50	0.0012	83	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
53	50	0.03	86	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
54	50	0.022	142	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
52	50	0.036	247	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
58	50	0.0094	108	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.042	89	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
55	50	0.016	157	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
53	50	0.029	154	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
51	50	0.038	119	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
59	50	0.0073	86	B. mallei GB5	B. mallei NCTC 10229	412022	Uninduced
56	51	0.017	215	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
67	51	0.0014	203	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
63	51	0.0035	161	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
57	51	0.014	71	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
57	51	0.015	144	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
56	51	0.018	82	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
53	51	0.036	123	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
52	51	0.042	121	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
54	51	0.031	111	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced

48	51	0.11	116	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
48	51	0.11	70	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
60	51	0.0074	74	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
55	51	0.026	73	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
54	51	0.031	178	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
54	51	0.032	181	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
53	51	0.033	136	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
52	51	0.042	136	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
56	51	0.017	236	B. pseudomallei 1126b	B. pseudomallei 668	320373	Uninduced
54	51	0.028	187	B. pseudomallei 1126b	B. pseudomallei 668	320373	Uninduced
57	51	0.016	50	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
52	51	0.048	74	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
52	51	0.046	117	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
68	51	0.0013	137	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
61	51	0.0056	172	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
59	51	0.0093	226	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
58	51	0.011	68	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
56	51	0.018	144	B. pseudomallei 1126B	B. pseudomallei 668	320373	Uninduced
57	50	0.01	165	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
51	50	0.043	219	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
68	50	0.00097	230	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
74	50	0.00022	251	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
57	50	0.01	251	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
55	50	0.018	223	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
54	50	0.022	82	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
54	50	0.023	211	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
51	50	0.041	112	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
51	50	0.044	82	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
55	50	0.019	220	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
48	50	0.086	325	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
50	50	0.058	62	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
50	50	0.054	330	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
49	50	0.065	136	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
49	50	0.079	259	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
88	50	0.0000082	255	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
51	50	0.041	163	B. thailandensis E264	B. thailandensis E264	271848	Uninduced
68	51	0.0013	162	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
69	51	0.00098	190	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
66	51	0.0017	103	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
52	51	0.047	182	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
49	51	0.095	148	B. vietnamensis FCO370	B. vietnamiensis G4	269482	Uninduced
45	51	0.23	197	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
53	51	0.038	163	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced
53	51	0.04	78	B. vietnamensis FCO369	B. vietnamiensis G4	269482	Uninduced

Sample	Spot Number	Comp.	Comp description	OLD TIGR	Possible Associations sea
10	11	Comp04	increased (i.e. Up) in 1126bI vs. 1126b		
10	111	Comp04	increased (i.e. Up) in 1126bI vs. 1126b		
10	111	Comp04	increased (i.e. Up) in 1126bI vs. 1126b		
10B	131	Comp18B	Unique to 10 vs 11		
10	135	Comp04	increased (i.e. Up) in 1126bI vs. 1126b		
10	154	Comp04B	increased (i.e. Up) in 1126bI vs. 1126b		
10	154	Comp04B	increased (i.e. Up) in 1126bI vs. 1126b		
10C	155	Comp04B	Increased in 10 vs 4		
10	159	Comp15	unique (Only) in GB8I vs. 1126I		
10B	174	Comp19B	Unique to 10 vs 12		
10A	207	Comp19B	Unique to 10 vs 12		
10	25	Comp4	unique (Only) in 1126bI vs. 1126b		
10	25	Comp04A	unique (Only) in 1126bI vs. 1126b		
10B	27	Comp4B	Increased in 10 vs 4		
10B	27	Comp4B	Increased in 10 vs 4		
10	47	Comp4	increased (Up) in 1126bI vs. 1126b		
10	47	Comp4	increased (Up) in 1126bI vs. 1126b		
10	62	Comp04A	increased (Up) in in 1126bI vs. 1126b		
10	67	Comp19A	Unique to 10 vs 12		
10	91	Comp18	unique (Only) in 1126bI vs. E254I		
11C	108	Comp5B	Increased in 11 vs 5		found on USAMRIID datab
11C	112	Comp20B	Unique to 11 vs 12		
11C	131	Comp05A	Unique to 11 vs 5		confirmed and corrected l
11B	137	Comp05B	Increased in 11 vs 5		confirmed and corrected l
11	139	Comp5B	Increased in 11 vs 5		comp20b (i.e., E256I vs.FC
11C	150	Comp05B	Increased in 11 vs 5		
11C	155	Comp05B	Increased in 11 vs 5		
11	16	Comp18	unique (Only) E264I vs. 1126bI		
11	161	Comp05	increased (Up) in E254I vs. E254		
11B	184	Comp05B	Increased in 11 vs 5		
11A	19	Comp05B	Increased in 11 vs 5		
11C	191	Comp18C	Unique to 11 vs 10		
11C	208	Comp18	Unique to E264I vs		
11B	218	Comp05B	Increased in 11 vs 5		
11B	218	Comp05B	Increased in 11 vs 5		
11	22	Comp5B	Increased in 11 vs 5		
11	22	Comp5B	Increased in 11 vs 5		
11C	240	Comp5B	Increased in 11 vs 5		
11	241	Comp5B	Increased in 11 vs 5		
11	241	Comp5B	Increased in 11 vs 5		
11	243	Comp18C	Unique to 11 vs 10		
11	47	Comp05B	Increased in 11 vs 5		
11	49	Comp05B	Increased in 11 vs 5		
11	57	Comp05B	Increased in 11 vs 5		
11C	72	Comp20B	Unique to 11 vs 12		
11	78	Comp18C	Unique to 11 vs 10		
11	78	Comp18C	Unique to 11 vs 10		
11	81	Comp05B	Increased in 11 vs 5		
11	90	Comp05B	Increased in 11 vs 5		
12	1				
12	101	Comp06B	Increased in 12 vs 6		
12C	117	Comp06A	Unique to 12 vs 6		
12C	154	Comp19C	Unique to 12 vs 10		
12	156	Comp06B	Increased in 12 vs 6		
12	160	Comp06B	Increased in 12 vs 6		
12	167	Comp06B	Increased in 12 vs 6		
12C	171	Comp17A	Unique to 12 vs 7		

12C	197	Comp17C	Unique to 12 vs 7	
12D	198	Comp17C	Unique to 12 vs 7	
12D	198	Comp17C	Unique to 12 vs 7	
12C	218	Comp19C	Unique to 12 vs 10	
12	22	Comp6B	Increased in 12 vs 6	
12	22	Comp6B	Increased in 12 vs 6	
NR	26	Comp17C	Unique to 12 vs 7	not recorded as having be
12	37	Comp06B	Increased in 12 vs 6	
12C	39	Comp06B	Increased in 12 vs 6	
12C	39	Comp06B	Increased in 12 vs 6	
12C	46	Comp06B	Increased in 12 vs 6	
12C	73	Comp06B	Increased in 12 vs 6	
12C	83	Comp06B	Increased in 12 vs 6	
12C	83	Comp06B	Increased in 12 vs 6	
12C	89	Comp06B	Increased in 12 vs 6	
12C	24	Comp06A	Unique in 12 vs 6	
12D	93	Comp06A	Unique in 12 vs 6	
12C	44	Comp06B	Increased in 12 vs 6	no spectrum but all other
6C	102	Comp06C	Unique to 6 vs 12	
6C	118	Comp25C	Unique to 6	
6C	118	Comp25C	Unique to 6	
6C	125	Comp26C	Unique to 6 vs 5	
6C	136	Comp23C	Unique to 6 vs 12	
NR	15	Comp03B	Increased in 9 vs 3	found by matching spot v
6	69			
6	69			
6	99	Comp06C	Unique to 6 vs 12	
7A	101	Comp01B	Increased in 7 vs 1	
7A	132	Comp17B	Unique to 7 vs 12	
7B	134	Comp17B	Unique to 7 vs 12	
7	142	Comp7B	Unique to 7 vs 8	
7B	153	Comp01B	Increased in 7 vs 1	
7B	153	Comp01B	Increased in 7 vs 1	
7B	153	Comp01B	Increased in 7 vs 1	
7A	142	Comp13B	Unique to 7 vs 8 and 9	
7A	167	Comp13B7	Unique to 7 vs 8 and 9	
7A	167	Comp13B7	Unique to 7 vs 8 and 9	
7A	167	Comp13B7	Unique to 7 vs 8 and 9	
7C	178	Comp16B7	Unique to 7 vs 11	
7C	179	Comp16B7	Unique to 7 vs 11	
7C	19	Comp17B	Unique to 7 vs 12	
7C	20	Comp08B	Unique to 7 vs 9	
7C	20	Comp08B	Unique to 7 vs 9	
7C	27	Comp15B	Unique to 7 vs 10	
7A	31	Comp17B	Unique to 7 vs 12	
7A	31	Comp17B	Unique to 7 vs 12	
7A	59	Comp13B7	Unique to 7 vs 8 and 9	
7C	63	Comp07B	Unique to 7 vs 8	
7C	69	Comp15B	Unique to 7 vs 10	
7C	9	Comp01A	Unique to 7 vs 1	
7A	95	Comp08B	Unique to 7 vs 9	
7	97	Comp01B	Increased in 7 vs 1	
7	97	Comp01B	Increased in 7 vs 1	
8	12			
8C	124	Comp02B	Increased in 8 vs 2	
8B	131	Comp09B	Unique to 8 vs 9	
8B	131	Comp09B	Unique to 8 vs 9	
8C	133	Comp02B	Increased in 8 vs 2	

8C	133	Comp02B	Increased in 8 vs 2	
8	136	Comp02	Increased in 8 vs 2	
8	136	Comp02	Increased in 8 vs 2	
8	141	Comp07C	Unique to 8 vs 7	
8	144	Comp09B	Unique to 8 vs 9	
8	16	Comp13	unique (Only) in GB6I vs. GB8I & GB5I	
8B	161	Comp02B	Increased in 8 vs 2	
8B	161	Comp02B	Increased in 8 vs 2	
8C	167	Comp2B	Increased in 8 vs 2	
8C	167	Comp2B	Increased in 8 vs 2	
8B	172	Comp02B	Unique to 8 vs 2	
8C	181	Comp02A	Increased in 8 vs 2	
8A	183	Comp13B8	Unique to 8 vs 7 and 9	
8A	191	Comp02A	Unique to 8 vs 2	
8A	191	Comp02A	Unique to 8 vs 2	
8B	206	Comp02B	Increased in 8 vs 2	
8C	29	Comp09B	Unique to 8 vs 9	unique to 8 vs 9 according
8	30	Comp13B8	Unique to GB6 induced vs GB8 induced and GB5 induced	
8B	318	Comp02B	Increased in 8 vs 2	
8B	53	Comp02B	Increased in 8 vs 2	
8B	53	Comp02B	Increased in 8 vs 2	
8	57	Comp2B	Increased in 8 vs 2	
8B	78	Comp7C	Unique to 8 vs 7	
9C	1	Comp13B9	Unique to GB5I vs GB8I and GB6I	
9B	119	Comp8C	Unique to 9 vs 7	
9	12	Comp09C	Unique to 9 vs 8	
9B	12	Comp09C	Unique to 9 vs 8	
9B	121	Comp03A	Unique to 9 vs 3	
9B	121	Comp03A	Unique to 9 vs 3	
9	162	Comp03B	Increased in 9 vs 3	
9	162	Comp03B	Increased in 9 vs 3	
9	164	Comp03B	Increased in 9 vs 3	
9C	165	Comp09C	Unique to 9 vs 8	
9C	165	Comp09C	Unique to 9 vs 8	
9C	165	Comp09C	Unique to 9 vs 8	
9C	182	Comp03A	Unique to 9 vs 3	
9A	183	Comp03B	Increased in 9 vs 3	
9A	279	Comp03B	Increased in 9 vs 3	
9C	294	Comp08C	Increased in 9 vs 3	
9C	294	Comp08C	Increased in 9 vs 3	
9B	3	Comp03B	Unique to 9 vs 7	
9C	48	Comp03A	Unique to 9 vs 3	
9	78	Comp3B	Increased in 9 vs 3	
1	11	Comp14	Unique to GB8 uninduced vs GB6 ur	chaperone protein DnaK (dnaK) {E
1	11	Comp14	Unique to GB8 uninduced vs GB6 ur	chaperone protein DnaK (dnaK) {E
1	11	Comp14	Unique to GB8 uninduced vs GB6 ur	chaperone protein DnaK (dnaK) {E
1	11	Comp14	Unique to GB8 uninduced vs GB6 ur	chaperone protein DnaK (dnaK) {E
1	11	Comp14	Unique to GB8 uninduced vs GB6 ur	chaperone protein DnaK (dnaK) {E
1A	73	Comp01C	Unique to 1 vs 7	arginine deiminase (arcA) [3.5.3.6]
1A	73	Comp01C	Unique to 1 vs 7	arginine deiminase (arcA) [3.5.3.6]
1A	73	Comp01C	Unique to 1 vs 7	arginine deiminase (arcA) [3.5.3.6]

1B	52	Comp10B	Unique to 1 vs 2	glutamine synthetase, type I (glnA
1B	52	Comp10B	Unique to 1 vs 2	glutamine synthetase, type I (glnA
1B	52	Comp10B	Unique to 1 vs 2	glutamine synthetase, type I (glnA
1B	52	Comp10B	Unique to 1 vs 2	glutamine synthetase, type I (glnA
1B	55	Comp22B	Unique to 1 vs 5	trigger factor (tig) [5.2.1.8] {Burkho
1B	55	Comp22B	Unique to 1 vs 5	trigger factor (tig) [5.2.1.8] {Burkho
1B	55	Comp22B	Unique to 1 vs 5	trigger factor (tig) [5.2.1.8] {Burkho
1B	55	Comp22B	Unique to 1 vs 5	trigger factor (tig) [5.2.1.8] {Burkho
1B	96	Comp10B	Unique to 1 vs 2	isocitrate dehydrogenase, NADP-c
1B	96	Comp10B	Unique to 1 vs 2	isocitrate dehydrogenase, NADP-c
1B	97	Comp23B	Unique to 1 vs 6	isovaleryl-CoA dehydrogenase (ivo
1B	97	Comp23B	Unique to 1 vs 6	isovaleryl-CoA dehydrogenase (ivo
1B	97	Comp23B	Unique to 1 vs 6	isovaleryl-CoA dehydrogenase (ivo
NR	112	Comp23A	Common to 1 and 6	acetylornithine aminotransferase (arg
NR	112	Comp23A	Common to 1 and 6	acetylornithine aminotransferase (arg
NR	112	Comp23A	Common to 1 and 6	acetylornithine aminotransferase (arg
NR	112	Comp23A	Common to 1 and 6	acetylornithine aminotransferase (arg
NR	112	Comp23A	Common to 1 and 6	acetylornithine aminotransferase (arg
1B	117	Comp21B	Unique to 1 vs 4	conserved hypothetical protein {Bu
1B	117	Comp21B	Unique to 1 vs 4	conserved hypothetical protein {Bu
1B	117	Comp21B	Unique to 1 vs 4	conserved hypothetical protein {Bu
1B	117	Comp21B	Unique to 1 vs 4	conserved hypothetical protein {Bu
1	131	Comp23	Unique (Only) to GB8 vs FCO369	translation elongation factor Ts (tsf) {
NR	161	Comp10A	Common to 1 and 2	spot 161 volume 23.925 fr
1B	166	Comp21B	Unique to 1 vs 4	ferredoxin--NADP reductase (fpr) [
1B	166	Comp21B	Unique to 1 vs 4	ferredoxin--NADP reductase (fpr) [
1B	175	Comp22B	Unique to 1 vs 5	antioxidant, AhpC/Tsa family {Burk
1B	175	Comp22B	Unique to 1 vs 5	antioxidant, AhpC/Tsa family {Burk
1B	231	Comp01C	Unique to 1 vs 7	phasin family protein {Burkholderia m
1C	8	Comp23B	Unique to 1 vs 6	polyribonucleotide nucleotidyltrans
1C	8	Comp23B	Unique to 1 vs 6	polyribonucleotide nucleotidyltrans
1C	26	Comp10B	Unique to 1 vs 2	acetyl-CoA carboxylase, biotin car
1B	27	Comp11B	Unique to 1 vs 3	prolyl-tRNA synthetase (proS) [6.1
1C	41	Comp01C	Unique to 1 vs 7	chaperonin, 60 kDa (groEL) {Burkh
1C	41	Comp01C	Unique to 1 vs 7	chaperonin, 60 kDa (groEL) {Burkh
1C	41	Comp01C	Unique to 1 vs 7	chaperonin, 60 kDa (groEL) {Burkh
1C	51	Comp11B	Unique to 1 vs 3	
1C	60	Comp01C	Unique to 1 vs 3	glutamine synthetase, type I (glnA) [6
1C	60	Comp01C	Unique to 1 vs 3	glutamine synthetase, type I (glnA) [6
1B	99	Comp23B	Unique to 1 vs 6	isocitrate dehydrogenase, NADP-c
1B	99	Comp23B	Unique to 1 vs 6	isocitrate dehydrogenase, NADP-c
1B	99	Comp23B	Unique to 1 vs 6	isocitrate dehydrogenase, NADP-c
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme
1C	111	Comp21	Unique to 1 vs 4	syringomycin biosynthesis enzyme

2A	140	Comp02C	Unique to 2 vs 8	translation elongation factor Tu (tu
2A	177	Comp10C	Unique to 2 vs 1	
2A	177	Comp10C	Unique to 2 vs 1	
2A	177	Comp10C	Unique to 2 vs 1	
2A	177	Comp10C	Unique to 2 vs 1	
2A		Comp02C	Unique to 2 vs 8	
2C	37	Comp10C	Unique to 2 vs 1	acetyl-CoA carboxylase, biotin car
2C	61	Comp02C	Unique to 2 vs 8	extracellular nuclease, putative {Bu
2C	61	Comp02C	Unique to 2 vs 8	extracellular nuclease, putative {Bu
2C	61	Comp02C	Unique to 2 vs 8	extracellular nuclease, putative {Bu
2C	102	Comp02C	Unique to 2 vs 8	glycerol kinase (glpK) [2.7.1.30] {B
2C	102	Comp02C	Unique to 2 vs 8	glycerol kinase (glpK) [2.7.1.30] {B
2C	102	Comp02C	Unique to 2 vs 8	glycerol kinase (glpK) [2.7.1.30] {B
2C	102	Comp02C	Unique to 2 vs 8	glycerol kinase (glpK) [2.7.1.30] {B
2	176	Comp14	Unique to GB6 uninduced vs GB8 ur	isocitrate dehydrogenase, NADP-depe
2	176	Comp14	Unique to GB6 uninduced vs GB8 ur	isocitrate dehydrogenase, NADP-depe
3C	296	Comp03C	Unique to 3 vs 9	
3C	9	Comp11C	Unique to GB5 uninduced vs GB8 ur	acetyl-CoA carboxylase, biotin car
3C	9	Comp11C	Unique to GB5 uninduced vs GB8 ur	acetyl-CoA carboxylase, biotin car
3C	9	Comp11C	Unique to GB5 uninduced vs GB8 ur	acetyl-CoA carboxylase, biotin car
3C	51	Comp03C	Unique to 3 vs 9	
3C	324	Comp03C	Unique to 3 vs 9	isocitrate dehydrogenase, NADP-c
3C	324	Comp03C	Unique to 3 vs 9	isocitrate dehydrogenase, NADP-c
3	330	Comp11	Unique to GB5 uninduced vs GB8 ur	carbohydrate porin, OprB family {E
3	330	Comp11	Unique to GB5 uninduced vs GB8 ur	carbohydrate porin, OprB family {E
3	338	Comp11	Unique to GB5 uninduced vs GB8 ur	syringomycin biosynthesis enzyme
3	338	Comp11	Unique to GB5 uninduced vs GB8 ur	syringomycin biosynthesis enzyme
3	338	Comp11	Unique to GB5 uninduced vs GB8 ur	syringomycin biosynthesis enzyme
3	339	Comp12	Unique to GB5 uninduced vs GB6 ur	rod shape-determining protein Mre
4C	40	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	119	Comp04C	Unique to 4 vs 10	
4C	120	Comp25B	Unique to 4 vs 6	

4C	134	Comp25B	Unique to 4 vs 6	electron transfer flavoprotein, alph
4C	144	Comp25B	Unique to 4 vs 6	conserved hypothetical protein {Bu
4C	182	Comp21C	Unique to 4 vs 1	-
4C	182	Comp21C	Unique to 4 vs 1	-
4C	182	Comp21C	Unique to 4 vs 1	-
4C	194	Comp24B	Unique to 4 vs 5	
4C	194	Comp24B	Unique to 4 vs 5	
4C	194	Comp24B	Unique to 4 vs 5	
4	208	Comp21C	Unique to 4 vs 1	
4	208	Comp21C	Unique to 4 vs 1	
4C	213	Comp21C	Unique to 4 vs 1	conserved hypothetical protein {Bu
4C	213	Comp21C	Unique to 4 vs 1	conserved hypothetical protein {Bu
4C	221	Comp21C	Unique to 4 vs 1	conserved hypothetical protein {Bu
4C	42	Comp04C	Unique to 4 vs 10	serine-type carboxypeptidase fami
4C	42	Comp04C	Unique to 4 vs 10	serine-type carboxypeptidase fami
4C	42	Comp04C	Unique to 4 vs 10	serine-type carboxypeptidase fami
4C	42	Comp04C	Unique to 4 vs 10	serine-type carboxypeptidase fami
4C	42	Comp04C	Unique to 4 vs 10	serine-type carboxypeptidase fami
5B		Comp05C	Unique to 5 vs 11	
5B		Comp05C	Unique to 5 vs 11	
5B	12	Comp24C	Unique to 5 vs 4	serine-type carboxypeptidase fami
5B		Comp22C	Unique to 5 vs 1	
5B	29	Comp05C	Unique to 5 vs 11	
5B	29	Comp05C	Unique to 5 vs 11	
5B	29	Comp05C	Unique to 5 vs 11	
5B	29	Comp05C	Unique to 5 vs 11	
5B	55	Comp26B	Unique to 5 vs 6	translation elongation factor Ts (ts
5B	63	Comp05C	Unique to 5 vs 11	conserved hypothetical protein {Bu
5B	71	Comp22C	Unique to 5 vs 1	3-oxoadipate CoA-succinyl transfe
5B	75	Comp22C	Unique to 5 vs 1	3-oxoadipate CoA-succinyl transfe
5B	90	Comp22C	Unique to 5 vs 1	universal stress protein family {Bu
5B	105	Comp05C	Unique to 5 vs 6	heat shock protein HtpG (htpG) {B
5B	105	Comp05C	Unique to 5 vs 6	heat shock protein HtpG (htpG) {B
5B	123	Comp05C	Unique to 5 vs 11	PspA/IM30 family protein {Burkhol
5B	134	Comp26B	Unique to 5 vs 6	ATP synthase F1, beta subunit (atpD)
5B	134	Comp26B	Unique to 5 vs 6	argininosuccinate synthase (argG) [6.3
6B		Comp23C	Unique to 6 vs 1	
6B		Comp23C	Unique to 6 vs 1	
6B		Comp23C	Unique to 6 vs 1	
6B		Comp23C	Unique to 6 vs 1	
6C	12	Comp25C	Unique to 6 vs. 4	COG0554: Glycerol kinase [Burkholde
6C	38	Comp25C	Unique to 6 vs. 4	malate dehydrogenase (mdh) [1.1.1.3
6C	64	Comp06C	Unique to 6 vs 12	NADH dehydrogenase I, C subunit
6C	64	Comp06C	Unique to 6 vs 12	NADH dehydrogenase I, C subunit